Ottawa

DH2020

carrefours/
intersections

conference | colloque **July 22-24 juillet**

workshops | ateliers **July 20, 21, 25 juillet**

Ottawa
DH2020 carrefours/ intersections
conference | colloque July 22-24 juillet
workshops | ateliers July 20, 21, 25 juillet

Digital Humanities 2020
Book of Abstracts

University of Ottawa and Carleton University

Ottawa, Canada
July 20-25, 2020

uOttawa

Carleton UNIVERSITY

List of Abbreviations

Please note that this list reflects the abbreviations used in the editorial matter of this volume, but individual submissions might use additional acronyms.

aaDH - Australasian Association for Digital Humanities
ACH - Association for Computers and the Humanities
ADHO - Alliance of Digital Humanities Associations
CCC - Conference Coordinating Committee
CO - Constituent Organization
COB - ADHO Constituent Organization Board
CSDH/SCHN - Canadian Society for Digital Humanities / Société canadienne des humanités numériques
DH - Digital Humanities
DHASA - Digital Humanities Association of Southern Africa
DS/CN
EADH - European Association for Digital Humanities
EB - ADHO Executive Board
JADH - Japanese Association for Digital Humanities
LO - DH Conference Local Organizers
PC - DH Conference Program Committee
RedHD - Red de Humanistas Digitales
SSHRC/CRSH - Social Sciences and Humanities Research Council of Canada - Conseil de recherches en sciences humaines
TADH - Taiwanese Association for Digital Humanities

DH2020 Organizing Committees

PROGRAM COMMITTEE

The program committee undertook a significant amount of work towards the DH2020 conference. The PC helped craft our call for papers, building on the proposed theme the local organizers submitted in their original application three years prior. Individual members read dozens of applications in order to vet and shape a strong peer-reviewed program. They worked collaboratively to ensure that applicants' work was carefully and fairly considered and provided us as program chairs with vital information to aid in the final selection of the conference program. We would like to extend to each of the program committee members our thanks for their service on this committee.

Rahul K. Gairola, aaDH
Hugh Craig, aaDH
Anelise Hanson-Shrout, ACH
Angel David Nieves, ACH
Laura Mandell, CenterNet
Elli Mylonas, CenterNet
Alison Hedley, CSDH/SCHN
Bill Turkel, CSDH/SCHN
Karen Calteaux, DHASA
Juan Steyn, DHASA
James O'Sullivan, EADH
Elena Spadini, EADH
Aurélien Berra, Humanistica
Michael Sinatra, Humanistica
Tomoji Tabata, JADH
Taizo Yamada, JADH
Eloy Caloca, RedHD
Ernesto Priani, RedHD
Chao-Lin Liu, TADH
Muh-Chyun Tang, TADH
Laura Estill, co-chair
Jennifer Guiliano, co-chair

LOCAL ORGANIZING COMMITTEE

Constance Crompton, University of Ottawa
Sarah Simpkin, University of Ottawa
Kevin Kee, University of Ottawa
Shawn Graham, Carleton University
Brian Greenspan, Carleton University
Jada Watson, University of Ottawa

Sponsors and Acknowledgements

SPONSORS

SSHRC Connection Grant (funds)
uOttawa (funds and in-kind)
Carleton (in-kind)
Gale Cengage (funds)

ACKNOWLEDGEMENTS

On behalf of the local organisers Constance Crompton would like to thank May Ning, the unflagging DH2020 conference coordinator and designer. The elegance of the book of abstracts and website, the successful logistics coordination (including between Conftool and the University of Ottawa's registration system) are thanks to her particular genius. We would like to thank Huma Zafar for adapting and launching Convalidator to meet the needs of this book of abstracts and ADHO's abstract XML archive. We would also like to thank the University of Ottawa conference services, Carleton University conference services, and Tourism Ottawa for the services that they provided the conference free of charge, and as the live conference was cancelled, with no hope of financial benefit. We would like to thank Audrée Dubuc, Richard Boudria, Lynn Rivard, and the University of Ottawa's Faculty of Arts financial and legal teams, not only for underwriting the conference, but for facilitating registration, contracts, tax compliance, and other conference-related financial services. Finally we would like to thank Laura Estill and Jennifer Guiliano for their grace as program chairs, their willingness to have local organizers share in the intellectual work of the conference, and for their vision for the themes and inclusivity of the in-person DH2020 conference.

Laura Estill and Jennifer Guiliano would like to thank 2018 program committee chair Élika Ortega for her assistance and counsel during the conference planning process. We would also like to express our gratitude for the work of the local organizers, Constance Crompton, Sarah Simpkin, Shawn Graham, Brian Greenspan, Jada Watson, and Kevin Kee.

Behind the scenes, over the last eighteen months, Constance, Shawn, Sarah, Brian, Kevin and Jada have been a joy to collaborate with. They have worked tirelessly to create what would have certainly been a positive conference experience and environment. Their commitment to a diverse, inclusive, and supportive event made our work as program chairs possible. They continue that labour now as they work through the logistics of conference cancellation.

The DH2020 program committee has also undertaken a significant amount of work. The PC helped craft our call for papers, building on the proposed theme the local organizers submitted in their original application three years prior. Individual members read dozens of applications in order to vet and shape a strong peer-reviewed program. They worked collaboratively to ensure that applicants' work was carefully and fairly considered and provided us as program chairs with vital information to aid in the final selection of the conference program. We would like to extend to each of the program committee members our thanks for their service on this committee.

Editorial Process

Edited by
Constance Crompton
Laura Estill
Jennifer Guiliano

DHConvalidator Service
Created by Christof Shöch
Mounted and updated by Huma Zafar, with support from John Simpson and Compute Canada

Digital Abstract Interface Design and Search and Sorting Function
May Ning

Design & Typesetting
May Ning

Proof-Reading
Laura Estill and Jennifer Guiliano

Copyright
The abstract text are available under Creative Commons licence CC-BY-4.0

Online abstract available at: https://dh2020.adho.org/abstracts/

A Note on the Text

The abstracts presented in this volume were written by the contributors. Because DH2020 is an interdisciplinary conference that draws together researchers from across fields and from different geographical regions, we did not request a unified citation style. In "A Short Note on Pointless Reference Formatting," Philip C. Bayeve estimates that "that the research community spends upwards of three million hours every year" reformatting references; we did not want to add to that burden.

~Constance Crompton, Laura Estill, and Jennifer Guiliano

Bayeve, Philip. "A Short Note on Pointless Reference Formatting." Journal of Scholarly Publishing 44.3 (2013): 283-88.

If you would like a hard copy of the this book of abstracts, we recommend The Printing House's print-on-demand service. You can find out more about their binding services at https://www.tph.ca/printing-services/book-printing and can send this PDF for printing through https://www.tphdirect.ca/quickdirect2/filesender/023

Note on COVID-19 Pandemic

DH2020 was scheduled to take place in Ottawa, Canada from 19-25 July 2020, but the global COVID-19 coronavirus pandemic, of course, changed everything. ADHO, in consultation with the DH2020 Local Organizers and Program Committee, cancelled the in-person gathering in order to safeguard individual and communal health. Local organizers, in concert with the Program Chairs, worked to cancel the event including the planned excursions for our DH community: a fun run between the Carleton University and University of Ottawa campuses along the 188-year old Rideau canal, a behind-the-scenes tour of Canada's National Science and Technology Museum's new collections and conservation centre, a dragon boat tour of one of the Rideau Canal locks with a stop at local rapids.

Via a decision of the ADHO Constituent Board and the Executive Committee, DH2020 was suspended with a new organizing committee convened to organize and lead the virtual conference under the mantle of DH2020v.

Letter from the LO and PC Chairs

On behalf of the local organizing team at Carleton University and the University of Ottawa as well as members of the DH2020 Program Committee spread throughout the world, we wish to extend our thanks to all the contributors to the DH2020 Ottawa Conference.

The international conference planning process runs roughly two years. It comprises an elaborate undertaking that relies on volunteer labour: from the initial bidding process completed by local hosts to the work of members of the program committee who draft the call for papers and select the program, to members of the digital humanities community who act as reviewers and awards committee adjudicators. It is a process that relies on generosity and collaboration. And, in the 2020 cycle, one that asked members of the digital humanities community to trust us as your local organizers and program committee. The open peer review process was an unqualified success; we were stunned by the positivity authors and reviewers espoused in the open review stages within a critical intellectual framework. We also learned important lessons that we will be passing on to future conference committees in our final report that will report on innovations related to the conference workflow, submission ontologies, and the challenges faced by the membership in participating in a conference process marred by a global pandemic. You'll glimpse some of that information in this book of abstracts.

We, the Ottawa local organizers and program committee chairs, want to take a moment to highlight for you the hopes that we brought to the conference planning process. We wanted every attendee to feel welcome in Ottawa and at the conference. That welcome was a decidedly Indigenous one that recognized that our very act of gathering was built on colonial legacies that disenfranchised Indigenous peoples from their lands, communities, and families. We wanted more than a simple land acknowledgement; we wanted attendees to see the vitality that we see articulated by Indigenous communities and scholars. We also sought to center the values of public humanities that start with humans and the humane, the counterpoint that our community brings to global discourses that centre tools and technologies. We encouraged open-ness and aimed to foster discussions of open access, open scholarship, and ultimately the open minds we hope are represented in our community. As a collective, we are ever more convinced that digital humanities should encourage inclusion rather than exclusion; intellectualism rather than anti-intellectualism; and kindness rather than harshness.

Our goal was that attendees would leave feeling invigorated and part of a community that should continually be seeking to embody the importance of working collectively, positively, and with generosity. As we watch the events of the last few months, whether those related to the pandemic of coronavirus or the global anti-Black violence that is decimating communities and families, we are even more convinced that digital humanists should be engaged with the questions of public humanities, partnership, and open data. We had hoped that those joining us in Ottawa would have joined ongoing conversations about the dissemination of scholarship both in-process and in its final form, issues of attribution and credit, and the relationship between private corporations and the cultural record. Even though we are not meeting together in Ottawa, we hope that this book of abstracts serves a resource for you until we meet again.

Constance Crompton
Laura Estill
Shawn Graham
Brian Greenspan
Jennifer Guiliano
Sarah Simpkin
Kevin Kee
Jada Watson

DH2020 at a Glance

Note: This data represents a static view of the conference as it was on July 1, 2020. Some submissions were accepted in a different format than originally submitted due to reviewer feedback or author preference. They are accounted for in the submission format originally accepted. Additionally, as a result of the COVID-19 pandemic, there were a number of elective withdrawals after March 1, 2020 that are accounted for in this data snapshot.

total submission submitted	625
total accepted	339
% acceptance	54%

By type:

Lightning Talks	
# submissions submitted	36
# submissions accepted	32
% accepted	89%

Panels	
# submissions submitted	30
# submissions accepted	24
% accepted	77.42%

Posters	
# submissions submitted	114
# submissions accepted	102
% accepted	89%

Forum	
# submissions submitted	22
# submissions accepted	17
% accepted	77.27%

Short Presentations	
# submissions submitted	224
# submissions accepted	155
% accepted	70.45%

Workshops	
# submissions submitted	22
# submissions accepted	17
% accepted	77.27%

Long Presentations	
# submissions submitted	170
# submissions accepted	128
% accepted	75.29%

DH2020 and the Open Review Process

As part of the conference focus on open data and open scholarship, the DH2020 conference chairs, in consultation with the local organizers and ADHO leadership, elected to utilize an open peer-review process.

In the open peer-review workflow, submissions were made visible to assigned reviewers and the program committee. This included the names and affiliations of all submitters. Likewise, the names of the reviewers were made available to the program committee, others who reviewed the same proposal after they completed their review, and the proposal's author(s) after the initial reviews were completed. The submissions, reviews, and responses were not however made available to the public. The reviews and responses themselves remain confidential.

The move to include an open-peer review process for DH2020 was not undertaken lightly. We recognize that open peer review could have placed scholars in precarious, untenured, or problematic situations in the awkward position of being asked to be critical of a submission by a more established scholar. It could have, likewise, brought out the worst parts of subconscious bias that plague academia. To address these concerns, we enabled reviewers to refuse to review any submission where they had concerns on the impact of reviewing for personal or professional reasons. Similarly, we allowed authors, through a private message to the PC Chairs, to share any concerns about who might be assigned to review. For the 2020 conference cycle, we assigned 875 individual reviewers to complete 3639 reviews on 615 unique submissions. Reviewers completed 93% of those assigned reviews. Most submissions received four to five reviews each. Details on how reviews were assigned are available via the Program Committee chair blog post, Not Your Average Conference.

The program committee would like to extend its thanks to the peer reviewers who volunteered and completed their work:

Peer Reviewers
(alphabetical by surname)

Alfie Abdul-Rahman
Caroline Abela
Hadeer Aboelnagah
Giuseppe Abrami
Raja Adal
Robyn Adams
Ayodeji Ayoola Adedara
Maria Jose Afanador-Llach
Eugenia Afinoguenova
G. Edzordzi Agbozo
Henning Agt-Rickauer
Modupe Roseline Agunbiade
Aiko Aida
Patrick Ajibade
Hiroyuki Akama
Sümeyye Akça
Ayodele James Akinola
Alyssa Alabassi
Sepideh Alassi
Eric Carlson Alexander
Mark Andrew Algee-Hewitt
Susanna Alles Torrent
Stacy Allison-Cassin
Lorenzo Allori
Roger Alsop
Ersin Altin
Seraphim Alvanides
Chiara Alzetta
Clifford Anderson
Deborah (Debbie) Anderson
Janet Anderson
Talea Anderson
Wendy Anderson
Tara Lee Andrews
Alessio Antonini
Alyssa Emily Arbuckle
Florentina Armaselu
Taylor Arnold
Eduard Arriaga
Mathieu Aubin
Purbasha Auddy
Alison L. Babeu
Gabriela Baeza Ventura
Petra Bago
Anne Baillot

David Bainbridge
James William Baker
Twila Bakker
Brian Ballentine
Adrien Barbaresi
Laure Barbot
Nicola Barbuti
Elton Barker
Tully Barnett
Brett Barney
Paul Barrett
José Francisco Barrón Tovar
Valentina Bartalesi
Caroline Bassett
Joe Bauer
Ryan Frederick Baumann
Melodee Beals
Valérie Beaudouin
Devin Becker
Kaspar Beelen
Emily Bell
Andrea Bellandi
Angela Bellia
Pascal Belouin
Francesca Benatti
Harmony Bench
Clayton John Benjamin
Patricia Bentley
Maria Berger
Susan Jean Bergeron
Helena Bermúdez Sabel
Aurélien Berra
Todd Berreth
Elisa Eileen Beshero-Bondar
Lee Bessette
Eliza Bettinger
Kyle Bickoff
Laura Biesiadecki
Yuri Bizzoni
Olin Bjork
Michael Black
Christopher William Blackwell
Abdelaziz Blilid
Akwasi Bosompem Boateng
Lidia Bocanegra Barbecho
Enrico Bocciolesi
Mary K. Bolin
Alexandra Bolintineanu
Bruno Bon
Elizabeth Leigh Bonds
Helen Katherine Bones
Hugo Bonin
Kyle Paul Booten
Alison Booth
José Borbinha
Leo Born
Zoe Borovsky
Gauvain Bourgne
Mike Charles Bowman
Andy Boyles Petersen
Eleni Bozia
Carmen Brando
Lyndsay Bratton
Aaron Brenner
Peter Broadwell
Mackenzie Brooks
Andrew S. Brown
Susan Brown
Ben Brumfield
Sara Brumfield
Virginia Brussa
Tatiana Bryant
Marco Büchler
Nick Budak
Andiswa Bukula
Anne Burdick
Patrick J. Burns
Elisabeth Burr
Florian Cafiero
Karen Calteaux
Blanca Calvo Figueras
Alexandra Gago Câmara
Kelsey Cameron
Jason Camlot
Alberto Campagnolo
Jean-Baptiste Camps
Lucía Cantamutto
Sheila Carey
Patricia Lynn Carlton
Bryan Carter
Priscila Ramos Carvalho
Madeleine Casad
Vittore Casarosa
Tommaso Caselli
Robert Casties
Paul Caton
Hugh Cayless
Vanessa Ceia
Javier Cha
Alix Chagué
Nadine Chambers
Ashley Marian Champagne
Erik Malcolm Champion
Rico Chapman
Olivier Charbonneau
Louis Chartrand
Emmanuel Château-Dutier
Chih-Ming Chen
Kuang-hua Chen
Song Chen
Yueh-hui Vanessa Chiang
Maria Teresa Chiaravalloti
Dawn Childress
Tanya Chiykowski-Rathke
Crystal Nicole Chokshi
Caitlin Christian-Lamb
Tyng-Ruey Chuang
Neil Chue Hong
Jon Chun
Fabio Ciotti
Paolo Ciuccarelli
Arianna Ciula
Dave Clark
Frédéric Clavert
Corey Clawson
Tanya Clement
Claire Clivaz
Zach Coble
Giovanni Colavizza
Mariona Coll Ardanuy
Cindy Conaway
Anne Cong-Huyen
Melanie Conroy
Kyle Conway
Charles M. Cooney
T.L. Cowan
Heidi Christina Craig
Hugh Craig
Gregory Crane
Constance Crompton
Brian Croxall

James Cummings
Richard Cunningham
Ana Cuquerella Jiménez-Díaz
Lyne Da Sylva
Peter Daengeli
Fraser James Dallachy
Julia Damerow
Nathan Daniels
Kevin Day
Aline JE Deicke
Angelo Mario Del Grosso
Gimena del Rio Riande
Francesca Dell Oro
Seth Denbo
Catherine DeRose
Caleb Derven
Renée Desjardins
Rachel Di Cresce
Francesca Di Donato
Alessandro Di Ludovico
Wout Dillen
Joseph DiNunzio
Maya Dodd
Melissa Dollman
Quinn Dombrowski
Antoine Doucet
J. Stephen Downie
Emily Drabinski
Keli Du
Tinghui Duan
Ryan Christopher Dubnicek
Michèle Duguay
Suzanne Dumouchel
Andrew Dunn
Alexander Dunst
Marten Düring
Jessica Dussault
Amy Earhart
Kai Eckert
Maciej Eder
Patrick Egan
Øyvind Eide
Frank Ekeberg
Moses Effiong Ekpenyong
Randa El-Khatib
Eiman Elgewely
Katherine Elkins
Tommaso Elli
Iain Emsley
Bill Endres
Joshua Alma Enslen
Matt Erlin
Natalia Ermolaev
Miguel Escobar Varela
Bacem Essam
Emily Esten
Laura Estill
Sally A. Everson
Hammou Fadili
Rachel Faerber-Ovaska
Michael Falk
Augustine Farinola
Antoine Fauchié
Katherine Mary Faull
Kurt E Fendt
Katrina Fenlon
Sylvia Fernandez
Ginestra Ferraro
Kirsty Fife
Frank Fischer
Franz Fischer
Kathleen Fitzpatrick
Julia Flanders
Jamie Folsom
Stefania Forlini
Christopher W. Forstall
Greta Franzini
Francesco Freddolini
Erik Fredner
Nancy Friedland
Heather Froehlich
Francesca Frontini
Antonio Fruttaldo
Jonathan Furner
Simon Gabay
Dinara Gagarina
Francisco Gago Jover
Rahul Krishna Gairola
Radhika Gajjala
Alex Galarza
Isabel Galina Russell
Ioana Galleron
Liliane Gallet-Blanchard
Anne Garcia-Fernandez
Susan Garfinkel
Amy Gay
Bernhard C. Geiger
Arjun Ghosh
Francesca Giannetti
Alexander Gil
Ben Gillespie
Roger Christopher Gillis
Julie Giovacchini
Jost Gippert
Paul Girard
Amy Larner Giroux
Tami Gitto-Kania
Evelyn Gius
Grant Glass
Tassie Gniady
Matthew Gold
Laura Gonzales
Paul Matthew Gooding
Daniel James Gorman Jr.
Rafał Ludwik Górski
Michael Gossett
Makoto Goto
Steven Gotzler
Amanda Starling Gould
Karen F. Gracy
Shawn Graham
Martin Grandjean
Katrina Grant
Steven Gray
Jennifer Grayburn
Brian Greenspan
Hannah Griggs
Natalia Grincheva
Elizabeth Grumbach
Max Grüntgens
Marie-France Guenette
Dalia Guerreiro
Jennifer Guiliano
Corinne Guimont
Chelsea Gunn
Allyssa Guzman
Carina Guzman
Stephanie Ayla Habicht
S.E. Hackney
Patience Haggin
Thomas Nikolaus Haider

Myung-Ja (MJ) Han
Matthew Nathan Hannah
Vanessa Hannesschlaeger
Aneliese Hanson Shrout
Charlie Harper
James Buffington Harr, III
Katherine H. Hart
Kevin Scott Hawkins
Alison Hedley
Serge Heiden
Patrick Helling
Rachel Hendery
Ulrike Henny-Krahmer
Amanda Henrichs
Leah Henrickson
Jason A. Heppler
Lisa Hermsen
J. Berenike Herrmann
Stefan Heßbrüggen-Walter
Maria Isabel Hidalgo Urbaneja
Devin Higgins
Tuomo Hiippala
Shawn Hill
Torsten Hiltmann
Uta Hinrichs
Radim Hladík
Hou Ieong Ho
Jessica Hoare
Elisabeth Hobisch
Tobias Hodel
Rik Hoekstra
Armin Hoenen
Martin Holmes
Timo Homburg
David Lowell Hoover
Jan Horstmann
Natalie Houston
Christian Howard-Sukhil
Eric Hoyt
Jieh Hsiang
Chijui Hu
Alexander Huber
Damien Garrett Huffer
Pim Huijnen
Solenn Huitric
Vilja Hulden
Alex Humphreys
Amy Hunsaker
Elizabeth Hunter
Thembelihle Hwalima
Christy Lynn Hyman
Iuliia V. Iashchenko
Fatiha Idmhand
Dean Irvine
Leif Isaksen
Jennifer Isasi
Iliana Damirovna Ismakaeva
Marcela Isuster
Lubomir Ivanov
Arun Jacob
Hannah Jacobs
Mathieu Jacomy
Diane Jakacki
Andrew Janco
Stefan Jänicke
Fotis Jannidis
Eva Marie Jansen
Bruce Janz
Najla Jarkas
Adam Jatowt
Thessa Jensen
Jacob Jett
Xiaoshuang Jia
Jialei Jiang
Antonio Jimenez-Mavillard
Ana Jofre
Daniel Johnson
Elizabeth-Anne Johnson
Phyllis Johnson
Carrie Johnston
John Jones
Michael Alastair Jones
Justin Joque
Kate Joranson
Anna Katerina Jordanous
Patrick Juola
Zoltan Kacsuk
Marilia Kaisar
Julie Kane
Maeve Kane
Frederic Kaplan
Jojo Karlin
Jesse P. Karlsberg
Andre Kåsen
Nabil Kashyap
Micki Kaufman
Lyonel Kaufmann
Akihiro Kawase
Nathan Kelber
Margaret Kelleher
Sarah Irene Brutton Kenderdine
Corinne Kennedy
Spencer Keralis
Katia Lida Kermanidis
Christopher Kermorvant
Chelsea Oei Kern
Carsten Kessler
Foaad Khosmood
Nobuhiko Kikuchi
Whitt Kilburn
Heejin Kim
Hoyeol Kim
Minhyoung Kim
Miki Kimura
JaShong King
Trip Kirkpatrick
Asanobu Kitamoto
Inna Kizhner
Rebecca Sutton Koeser
Mikko Kristian Koho
Leonard Konle
Marijn Koolen
Michael J. Kramer
Florian Kräutli
Steven Krauwer
William Kretzschmar
Jörn Kreutel
Ross Deans Kristensen-McLachlan
Bärbel Kröger
Tetsuji Kuboyama
Ritesh Kumar
Ashanka Kumari
Kim Lacey
John Ladd
Leo Lahti
Maurizio Lana
Anouk Lang
Alison Langmead
David Lassner

Chiara Latronico
John Laudun
Nicola Laurent
John Lavagnino
Matthew Lavin
Alexei Lavrentiev
Olivier Le Deuff
Ryan Kyle Leach
Zoe Genevieve LeBlanc
Christophe Leblay
Claire S. Lee
Yoo Young Lee
Lauriane Lemasson
Piroska Lendvai
Maxim Lengo
Sharon M. Leon
Bin Li
Jiwen Liang
Amanda Marie Licastro
Raquel Liceras-Garrido
Kristen Lillvis
Nung-yao Lin
Shane Lin
Thea Lindquist
Margaret Linley
Thomas Lippincott
Andrew Lison
Chao-Lin Liu
Pietro Maria Liuzzo
Chester Ronald Liwosz
Brandon Locke
Steven Lubar
Ana Lucic
Sean Luyk
Rongqian Ma
Rooweither Mabuya
James MacKenzie
Prathmesh Rajeshbhai Madhu
Justin Madron
Mary-Margaret Mahoney
Chao Maina
Eetu Mäkelä
Enrique Mallen
Andrea Malone
Andrew Mancuso
Laura Mandell
Molly Mann
Norma Aida Manzanera Silva
Kristen Mapes
Rennie Mapp
Sophie Marcotte
Itay Marienberg-Millikowsky
Jani Marjanen
Kim Martin
Merisa Ariel Martinez
Bruno Emanuel Martins
Maciej Maryl
Joan Mas
Austin Mason
Ingrid Mason
Brigitte Mathiak
Lindsay Mattock
Brian Matzke
Aaron Mauro
David Maus
Katherine McDonough
John McEwan
Shane Adam McGarry
Emily McGinn
Jim McGrath
Katherine McLeod
Margot Lise Mellet
Ana Luísa Melro
Sarah Melton
Luis Meneses
Orr Menirom
Nirmala Menon
Megan Meredith-Lobay
Albert Meroño Peñuela
Don Merritt
Adam Mertel
Thomas C. Messerli
Daniele Metilli
Samuel Mickelson
Sarah Middle
Rachel Clare Midura
Jodi Mikesell
George K. Mikros
Ben Miller
Ian Milligan
Matthew Milner
David Mimno
Naveen Zehra Minai
Dennis Mischke
So Miyagawa
Maki Miyake
Jarah Moesch
Neeraj Mogla
M. Willis Monroe
Suzanne Alayne Moody
Gabriela Margarita Morales Medina
Ashley Caranto Morford
Paige Morgan
Suzanne Mpouli
Sanjin Muftic
Rikk Mulligan
Trevor Munoz
Rebecca Munson
Hajime Murai
Krista Alexandra Murchison
Cal Murgu
Maggie Murphy
Orla Murphy
Elli Mylonas
Kiyonori Nagasaki
Yoshikatsu Nagata
Pinkey Nainwani
Narayanamoorthy Nanditha
John Nerbonne
Claes Neuefeind
Gregory Trent Newton
Emmanuel Ngue Um
Annie Laurie Nichols
Florian Niebling
Kristoffer Nielbo
Angel David Nieves
Chifumi Nishioka
Nichole Misako Nomura
Sarah Noonan
Jack Norton
Krzysztof Nowak
Julianne Nyhan
Daniel O'Donnell
James O'Sullivan
James O'Sullivan
Alexander OConnor
Fabian Offert
Jessica Ogden
Toshinobu Ogiso
Kazushi Ohya

Felix Bayode Oke
Nozomu Okuda
Ovio Viorel Olaru
Jenn Olive
Tunde Olusola Ope-Davies
Christian-Emil Ore
Espen Ore
Boris V. Orekhov
Peter Organisciak
Élika Ortega
Stephen Gregory Osadetz
Jessica Otis
Max Overbeck
Trevor Owens
Burcak Ozludil
Kevin Page
Émilie Pagé-Perron
Albert Palacios
Gregory Palermo
Costas Papadopoulos
Georgios Papaioannou
Emilie Paquin
Mylene Pardoen
Elizabeth Parke
Jessica Parr
Nikolaus Parulian
Bill Pascoe
Alicia Peaker
Jérémie Pelletier-Gagnon
Miriam Peña-Pimentel
Yi-Fan Peng
Andrea Penso
Arthur Perret
Anna Perricci
Chloe Perry
Mark Edwin Peterson
Gesina A. Phillips
Davide Picca
Steffen Pielström
José Pino-Díaz
Michael Piotrowski
Carrie Pirmann
Petr Plechac
Marisa Plumb
Marcus Pöckelmann
Thierry Poibeau
Caitlin Pollock
Julia G. Polyck-O'Neill
Liviu Pop
Adam Porter
J.D. Porter
Monika Porwol
Nadezhda Povroznik
Gila Prebor
Todd Presner
Marc Priewe
Ellen Prokop
Catherine Psilakis
Marie Anna Puren
William Quinn
Milena Radzikowska
Valiur Rahaman
Jas Rault
Padmini Ray Murray
Cecily Colleen Raynor
Stacy Reardon
Simone Rebora
Gethin Rees
Jonathan Reeve
Amanda E. Regan
Georg Rehm
Allen Renear
Daniel Riaño Rufilanchas
Leah Richardson
Antonio Rico-Sulayes
Allen Beye Riddell
Mia Ridge
Jenna Rinalducci
Hannah Ringler
Roopika Risam
Zak Risha
António Rito Silva
Brad Rittenhouse
Jörg Ritter
Léon Robichaud
Peter Robinson
Yannick Rochat
Geoffrey Martin Rockwell
Dario Rodighiero
Glenn Roe
Torsten Roeder
Roxana-Sînziana Rogobete
Antonio Rojas Castro
Laurent Romary
C.A. Romein
Dolores Romero-López
Jan Rörden
Roberto Rosselli Del Turco
Paolo Rossini
Atanassova Rossitza
Michael Roth
Charlotte Roueché
Jesse Rouse
Paulina Rousseau
Vasileios Routsis
Dibyadyuti Roy
Andrew Ruis
Pablo Ruiz Fabo
Norman Rusin
Amanda Rust
Arjun Sabharwal
Shouji Sakamoto
Diego Andres Salcedo
Gabriele Salciute Civiliene
Enrica Salvatori
Marta Samokishyn
Ashley Sanders Garcia
Dave Sandor
Lidia Santarelli
Amelia Sanz
Alexandra Sarkozy
Michael Satlow
Danica Savonick
Jacques Savoy
Gamze Saygi
stefania scagliola
Andrea Scharnhorst
Walter Scheirer
Maximilian Schich
Philip Schildkamp
Benjamin M Schmidt
Desmond Schmidt
Sara A. Schmidt
Thomas Schmidt
Stefan Schmunk
Rachel Schnepper
Christof Schöch
Walter Scholger
Christoph Schommer
Daniela Monika Schulz
Mark Schwindt

Mehdy Sedaghat Payam
Ana Sedeno-Valdellos
Tatsuki Sekino
Kylie Seltzer
Anindita Basu Sempere
Donnie Sendelbach
Luigi Serra
Anusharani Sewchurran
Wenyi Shang
Ryan Benjamin Shaw
Ayako Shibutani
Masahiro Shimoda
Mark Shuttleworth
A.R. Siders
Lynne Siemens
Raymond George Siemens
Sara Sikes
Sarah Simpkin
Matthew Sims
Michael Sinatra
Stéfan Sinclair
Gurpreet Singh
Bobby Smiley
Jolene L. U. Smith
Pamela H. Smith
James Smithies
Lisa M. Snyder
Yuting Song
Elena Spadini
Corey Sparks
C. M. Sperberg-McQueen
Rachele Sprugnoli
Joseph Ssebulime
Peter Stadler
Rombert Stapel
Rachel Lynne Starry
Samara Hayley Steele
Zacharias Steyn
H. Siegfried Stiehl
Peter Anthony Stokes
Oliver Streiter
libi striegl
Dominique Stutzmann
Chris Alen Sula
Dheepa Sundaram
Apiwat Suyabodha
Chikahiko Suzuki
Elizabeth Anne Swanstrom
Max J. Symuleski
Victoria Szabo
Yuta Takahashi
Timothy Tambassi
Raluca A. Tanasescu
Muh-Chyun Tang
Timothy Tangherlini
Felicity Tayler
Jennifer Whitmer Taylor
Orion Teal
Melissa Terras
Tyrica Terry Kapral
Elisa Tersigni
Jeffrey Tharsen
Ruck Thawonmas
Laure Thompson
Christopher James Thomson
Sabine Thuillier
Lauren Tilton
Adam Tindale
Irene Tobón Restrepo
Mikko Tolonen
Toru Tomabechi
Gaelle Tonnellier
Rosa María Torres-Medina
Maurizio Toscano
Charles Travis
Peer Trilcke
Thorsten Trippel
Richard Tzong-Han Tsai
Lik Hang Tsui
Aaron Tucker
Serdar Tuncer
Charlotte Tupman
Nicolas Turenne
William Turkel
Ece Turnator
Robert Tweheyo
Angeliki Tzouganatou
Ted Underwood
John Merritt Unsworth
Anagha Uppal
Tolga Uslu
Victoria Valdes
Vanessa Valdés
Marieke van Erp
Arjan van Hessen
Victoria Anne Van Hyning
Peter Van Kranenburg
Ingeborg van Vugt
Menno van Zaanen
Joris J. Van Zundert
Tomas Vancisin
Roberto Vargas
Demival Vasques Filho
Annette Vee
Athanasios Velios
John T. Venecek
Pierre Vernus
Cristina Vertan
Vander Viana
Gabriel Viehhauser
Raffaele Viglianti
Francois Vignale
Carolina Villarroel
Lorella Viola
Amanda Wyatt Visconti
Fabio Vitali
Thorsten Vitt
Georg Vogeler
Armin Volkmann
Andrey Volodin
Alexander von Lünen
Jessica Wagner Webster
Katherine Walden
Rebekah Walker
Tomasz Walkowiak
Katherine L. Walter
Jiayun Wang
Lihua Wang
Sean Wang
Jennifer Ward
Claire Warwick
Andrew Watkins
Jada Emily Watson
Sean Weidman
Scott B. Weingart
Kristen Weischedel
Naomi Amelia Stewart Wells
Henry Alexander Wermer-Colan
Leon Casimir Wessels
Melvin Wevers

Zachary N Whalen
David Stewart White
Lars Wieneke
Olivia Wikle
Helene C. Williams
Isaac Maximilian Williams
Patrick Williams
Elizabeth Williamson
Julius Wilm
Amanda Wilson Bergado
Florian Windhager
Caroline Winter
Dirk Wintergrün
Avery Jacob Wiscomb
Christian Wittern
Mark Wolff
Lia Wolock
Jewon Woo
Glen Worthey
Ulrike Wuttke
Lu Xiao
Mitsuyoshi Yabe
Taizo Yamada
Sean A. Yeager
Xin Yin
Michael Miller Yoder
Setsuko Yokoyama
Gerben Zaagsma Zaagsma
Vika Zafrin
Joel Zapata
Ewa Zegler-Poleska
Jean-Daniel Zeller
Marcia Zeng
Kalliopi Zervanou
Guangwei Zhang
Liyuan Zhang
Mariana Ziku
Amelie Zöllner-Weber
Mariana Zorkina
Chiara Zuanni
Natasha Zwarich
Benjamin Zweig

Calls for Proposals (CFP)

We are proud to include the calls for proposals for this year's conference in order to promote transparency and to maintain a record of the conference. The DH2020 submission process represented a number of innovations including new presentation formats, updated word counts for proposals, new keyword ontologies, and a revised conference timeline to accommodate open peer review. This year, we were also pleased to translate the CFP into an indigenous language, Anishinabeg/Algonquin, in order to reflect the conference themes.

CFP: Anishinabeg/Algonquin

Anishinaabe waasigani-gikendaasowinan 2020 : "Nakwemoning"

Anishinaabe waasigani-gikendaasowinan wekwiinojig, Alliance of Digital Humanities Organizations (ADHO), gidandomigoom wii biijinizha'ameg waa dazhindamowegwen enji-maawanjidiyang, Ode'imini-giizis 20 – 25, 2020 iwedi Carleton Gichigabegikendaasowigamigong gaye Odaawaa Gichigabegikendaasowigamigong enji teg Odaawaa odenang, Zhaaganaashikiing.

Gidandomigoom wii biijinizha'ameg: Anishinaabemong, Zhaaganaashimong, Wemitigoozhimong, Maandaashinaabemong, Aanimaamong, Miishaakiganimong

Biijinizha'amok waa dazhindamowegwen: https://www.conftool.pro/dh2020/

Booch igo daa izhichigaade: Binaakwii-giizis 22, 2019 11:59 PM PST
Giwii wiindamaagoom biindigazhigoyeg: Onaabani-giizis 1, 2020
Gaa inaakonigaade apii daa izhiwebag maawajidiwin

Jibwaa-maawanjidiwin enji-anokiiyeg gaye naadamaagewinan: Ode'imini-giizis 20-21, 2020
Maawanjidiwin: Ode'imini-giizis 22-24, 2020
ADHO nibwaachiwewinan gaye ishkwaa-maawanjidiwin enji-anokiiyeg: Ode'imini-giizis 25, 2020

Apii dash ge maawanjidiyang miinawaa biboonagag 2020 gakina wii maamawi-dazhindamang "Nakwemoning," enji-maawandooseg miikanan. Memwech igo indandawendaamin wii mookinameg gegoo ge dazhindameg dash: Anishinaabe nandagikendaasowinan; maamawi-awiyag Anishinaabe waasigani-gikendaasowinan; gaye wenipazh gikendaasowin bagidendaagwag. Gidandomigoom gakina dash ge anokiiyeg wii nandagikendameg anoonji-anokiiwinan Anishinaabe waasigani-gikendaasowining.

Gidaa inaakonaanaawaa bezhigwan:

- Agwaakwa`iganan (Gidozhibii'aanaawaa zhego 250 – 500 ikidowinan gakaam waa dazhindaagwag)
- Dadaatabowewinan (5 dibaa'iganensan; Gidozhibii'aanaawaa zhego 250 – 500 ikidowinan gakaam waa dazhindaagwag)
- Noomag dazhindaagwadoon (10 dibaa'iganensan; Gidozhibii'aanaawaa zhego 250 – 500 ikidowinan gakaam waa dazhindaagwag)
- Ginwenzh dazhindagwadoon (20 dibaa'iganensan; Gidozhibii'aanaawaa zhego 750 – 1000 ikidowinan gakaam waa dazhindaagwag)
- Inaakonigaazowag wii dazhindamowaad gegoon (90 dibaa'iganensan; Gidozhibii'aanaawaa

250 ikidowinan gakaam waa dazhindaagwag gaye zhego 300-500 ikidowinan waa maamodazhindaagwag)
- Maamawi-dazhindamang (90 dibaa'iganensan; Gidozhibii'aanaawaa 500 ikidowinan gakaam waa dazhindaagwag)
- Jibwaa-maawanjidiwin enji-anokiiyeg gaye naadamaagewinan (2 dibaa'iganan maage 4 dibaa'iganan; 250 ikidowinan waa maamodazhindaagwag gaye ozhibii'igaadeg naadamaagewinan gemaa gikinoo'amaagewinan)

Gakina ozhibii'igaade daa aawendaagwag gaa enji onzikaag. Gego agindangegon inootaagewinan apii agindameg ikidowinan. Daga agindamok gaa inaakonigaadeg endaso dinowang waa biijinizha'igaadeg. SIG wekwiinojig wii maawanjidiwag jibwaa maamawi-maawanjidiyang wii naagadawaabanjigewaad gaye. Daga nandotawig SIG wekwiinojig jibwaa biijinizha'iyaang gegoo SIG naabinootaagwag.

DH2020 inaakonaawaa gakina wenaagadawaabanjigejig gaye begidinigejig wii gikenindiyeg. Gakina inaakonigaadeg dash gikendaagwag eta gishpin wenaagadawaabanjigejig, begidinigejig, maage wenaakonigejig aawiyeg.

Mii dash wii ezhi inaakonamaang:
- Gakina nitaa-ozhisijigaade gaye nitaa-nisidawendaagwad. (20%)
- Gakina dash nitaa-nandagikendaagwad, gidazhindaanaawaa gaa ezhi nandagikendameg igo gaye waabanda'iweyeg izhiwebag noongom ge anokaadameg. (30%)
- Aawendaagwad ge izhichigeyeg "Nakwemoning": Anishinaabe nandagikendaasowinan; maamawi-awiyag Anishinaabe waasigani-gikendaasowinan; gaye wenipazh gikendaasowin bagidendaagwag. (10%)
- Gakina igo dash waa dazhindamowegwen daa wenipazh nisidawendaagwag gaye daa weweni waawiindamaageyeg. (20%)
- Gakina dash daa aabajichigaade gaye chi-inendaagwag memindaage gishpin anokaadameg Anishinaabe waasigani-gikendaasowinan. (20%)

Anishinaabe waasigani-gikendaasowinan wekwiinojig, The Alliance of Digital Humanities Organizations (ADHO), aapiji chipiitendaanaawaa ayaagwaamiziwin, manaadenidiwin, gaye naagadanenidiwin enji maawanjidiyang ge gakina daa ganoonidiyang naadamaadiyang wii gikendaasoyang anoonji. Gakina wii bi izhaawag daa mino-doodamowaad gaa ezhi dibaakonigaadeg. Wenaagadawaabanjigejig daa izhiwebiziwag ge ozhibii'igaadeg omaa.

Zhooniyaa daa daapinigaade iidog gishpin oshki-anokiiyeg maage eshkam gikinoo'amaagoziyeg

Anishinaabe waasigani-gikendaasowinan wekwiinojig wii miizhigoyeg zhooniyaa eniwek gishpin oshki-anokiiyeg maage eshkam gikinoo'amaagoziyeg gaye izhichigeyeg maawanjidi'ing. Gakina gidaa doojigem wii izhiwidooyaang iwedi ADHO waasigani-mazinaa'iganing wiibaj 2020: http://www.adho.org. Mii igo eta zhooniyaa daa aabajichigaadeg gishpin agwaakwa`iganikeyeg maage dazhindameg gegoo.

Gidaa gagwejimaawaag DH2020 wegimaawijig gishpin gagwedweyeg: Laura Estill miinawaa Jennifer Guiliano izhinikaazowaad pc2020@adho.org

Translation by Michael Zimmerman, Jr.

CFP: English

Digital Humanities 2020 : "Carrefours/Intersections"

The Alliance of Digital Humanities Organizations (ADHO) invites submission of proposals for its annual conference, July 20-25, 2020 at Carleton University and the University of Ottawa in Ottawa, Canada. Submissions will be accepted in Algonquin, English, French, German, Italian, and Spanish.

CFP: Algonquin, English, French, Spanish, German, Italian

Submit a Proposal: https://www.conftool.pro/dh2020/

Submission Guidelines

Deadline for Submissions: October 22, 2019 11:59 PM EDT (updated 10/14)

Notification of Acceptances: March 1, 2020

Scheduled Conference Dates
Pre-Conference Workshops and Tutorials: July 20-21, 2020
Conference: July 22-24, 2020
ADHO Excursions and Post-Conference Workshops: July 25, 2020

The theme of the 2020 conference is "Carrefours/Intersections," a place where roads or streets meet. We specifically invite proposals that relate to our sub-disciplinary conference interests: First Nations, Native American, and Indigenous Studies; public digital humanities; and the open data movement. We welcome all who identify themselves as working in the broad variety of disciplines, methodologies, and pedagogies that the digital humanities encompasses.

Submissions may include:
Posters (abstract 250-500 words)

- Lightning talks (5 minutes; abstract 200-250 words)
- Short presentations (10 minutes; abstract 250-500 words)
- Long presentations (20 minutes; abstract 750-1000 words)
- Panels (90 minutes; abstract 250 words + 300-500 word overview)
- Forums (90 minutes; abstract 500 words)
- Pre-conference workshops and tutorials (2 hours or 4 hours; 250-word overview plus syllabus and/or relevant existing tutorials)

All proposals should include relevant citations to sources in the appropriate literatures. Citations are not to be included in the word count. Please consult the detailed requirements for each submission type as well as submission guidelines. SIG groups will be holding topical pre-conference meetings that will undergo separate review via SIG organizers. Please contact SIG organizers before submitting a workshop or tutorial proposal on a duplicative SIG topic.
Representing the conference theme, DH2020 will use the open peer review process, which requires submitter and reviewer identities to be disclosed. Reviews will remain confidential to the submitter, reviewers, and the members of the Program Committee.

Submissions will be evaluated based on:

- Overall organization and clarity of proposed submission (20%)
- Explicit engagement with relevant scholarship, with references and justifications displaying knowledge of the current state of appropriate fields (30%)
- Thematic relevance to "carrefours/intersection"; Native American, Indigenous, and First Nations Studies; public digital humanities; or the open data movement (10%)
- Clear theoretical, methodological, or pedagogical framework and explicit statement of purpose (20%)
- Applicability, significance, and value of the theoretical, methodological, and/or practical contribution to the digital humanities generally (20%)

The Alliance of Digital Humanities Organizations (ADHO) is dedicated to creating a safe, respectful, and collegial conference environment for the benefit of everyone who attends and for the advancement of research and scholarship in fields supported by our constituent organizations. All authors and attendees are expected to abide by the Code of Conduct. Reviewers will be asked to follow the guidelines for reviewers.

BURSARIES FOR EARLY-CAREER AND EMERGING SCHOLARS

The Alliance of Digital Humanities Organizations will offer a limited number of bursaries for students and early-career scholars presenting at the conference. Bursary information is available on our Bursaries page.

Questions related to this CFP should be directed to the DH2020 Program Chairs, Laura Estill and Jennifer Guiliano at pc2020@adho.org

CFP: Español

Digital Humanities 2020 (DH2020): "Carrefours/Intersections"

La Alianza de Organizaciones de Humanidades Digitales (ADHO) invita a la presentación de propuestas para su conferencia anual que se llevará a cabo del 20-25 de julio del 2020 en la Universidad de Carleton y la Universidad de Ottawa en Ottawa, Canadá. Se aceptan propuestas en algonquino, inglés, francés, alemán, italiano y español.

Convocatoria: algonquino, inglés, francés, español, alemán, italiano

Presenta una propuesta: https://www.conftool.pro/dh2020/

Fecha límite para las presentaciones: 22 de octubre de 2019 a las 11:59 p.m. EDT
Notificación de aceptaciones: 1 de marzo de 2020

Fechas programadas
Talleres y tutoriales previos a la conferencia: 20-21 de julio de 2020
Conferencia: 22-24 de julio de 2020
Excursiones y talleres de ADHO posteriores a la conferencia: 25 de julio de 2020

El tema de la conferencia de 2020 es "Intersecciones", un lugar donde las carreteras o calles se encuentran. De manera específica, invitamos propuestas que se relacionen con los intereses sub-disciplinarios de nuestra conferencia: estudios de primeras naciones, nativos americanos e indígenas; humanidades digitales públicas; y el movimiento de datos abiertos.

Damos la bienvenida a todos los que se identifican dentro de la amplia variedad de disciplinas, metodologías y pedagogías que abarcan las humanidades digitales.
Las presentaciones pueden ser:

- Posters (resumen de 200-500 palabras)
- Charlas relámpago (5 minutos; resumen de 200-250 palabras)
- Presentaciones breves (10 minutos; resumen de 250-500 palabras)
- Presentaciones largas (20 minutos: resumen de 750-1000 palabras)
- Paneles (90 minutos; resumen de 250 palabras + un resumen general de 300-500 palabras)
- Foros (90 minutos: resumen de 500 palabras)
- Talleres y tutoriales previos al congreso (2 o 4 horas: resumen general de 250 palabras más programa y/o tutoriales relevantes existentes)

Todas las propuestas deben incluir citas y fuentes relevantes en la literatura apropiada. Las referencias no cuentan para los límites de palabras. Favor de consultar los requisitos detallados para cada tipo de presentación, así como para los modelos de presentación. Los Grupos de Interés Especial (SIG, por sus siglas en inglés) celebrarán reuniones temáticas previas a la conferencia que se someterán a una revisión por separado a través de los organizadores de los SIG. Favor de contactar a los organizadores de los SIG antes de enviar una propuesta de taller o tutorial sobre un tema en concordancia con los temas de los SIG.

Representado el tema de la conferencia, DH2020 utilizará el proceso abierto de revisión entre pares, que requiere que las identidades del remitente y del revisor se revelen entre sí. Las revisiones se mantendrán confidenciales para el remitente, los revisores y los miembros del

Comité del Programa.

Las propuestas serán evaluadas en base a:

- Organización y claridad general de la propuesta (20%)
- Compromiso explícito con temas académicos relevantes, que incluyan referencias y justificaciones que muestren el conocimiento del estado actual de los campos en cuestión (30%)
- Relevancia temática para "Intersecciones": Estudios de nativos americanos, indígenas y primeras naciones; humanidades digitales públicas; o el movimiento de datos abiertos (10%)
- Marco teórico, metodológico o pedagógico claro y declaración explícita de propósito (20%)
- Aplicabilidad, importancia y valor de la contribución teórica, metodológica y/o práctica a las humanidades digitales en general (20%)

La Alianza de Organizaciones de Humanidades Digitales (ADHO) está dedicada a crear un entorno seguro, respetuoso y colegiado para el congreso, en beneficio de todos los asistentes, y para el avance de la investigación y el trabajo académico en los campos que abarcan las organizaciones constituyentes. Se espera que todos los autores y asistentes se atengan al Código de Conducta. Se le pedirá a los revisores que sigan las pautas para los revisores

BECAS PARA ACADÉMICOS EMERGENTES Y EN ETAPAS TEMPRANAS DE SUS CARRERAS
La Alianza de Organizaciones de Humanidades Digitales ofrecerá un número limitado de becas para académicos en etapas tempranas de sus carreras que hayan sido aceptados para presentar en la conferencia. Los lineamientos para obtener estos beneficios aparecerán en el pagina de becas. Las becas solamente serán otorgadas para pósters y presentaciones.
Cualquier pregunta relacionada con esta convocatoria debe ser dirigida a las presidentas del programa DH2020, Laura Estill and Jennifer Guiliano, pc2020@adho.org.

Tradotto da Fabiola Delfín por RedHD

CFP: Deutsch

Digitale Geisteswissenschaften 2020: "Kreuzungen"

Die Alliance of Digital Humanities Organizations (ADHO) lädt zur Einreichung von Vorschlägen zu ihrer jährlichen Konferenz (20 -25 Juli 2020 in Ottawa) ein. Die Konferenz wird von der Carleton University und der University of Ottawa organisiert. Beiträge können in Algonkin, Englisch, Französisch, Spanisch, Deutsch und Italienisch verfasst werden.

Einreichung von Vorschlägen: Algonkin, Englisch, Französisch, Spanisch, Deutsch, Italienisch

Vorschlag-Einreichung: https://www.conftool.pro/dh2020/

Frist für die Einreichung von Vorschlägen: 22. Oktober 2019, 11:59 PM PST
Benachrichtigung über die Annahme: 1. März, 2020

Geplante Konferenz-Termine:
Workshops und Tutorien vor der Konferenz : 20-21 Juli 2020
Konferenz: Juli 22-24, 2020 /22-24 Juli 2020
ADHO Ausflüge und Workshops nach der Konferenz: 25. Juli 2020

Das Konferenz-Thema im Jahr 2020 is "Kreuzungen", ein Ort wo Wege oder Straßen sich treffen. Wir laden ausdrücklich zur Einreichung von Vorschlägen ein, die mit unseren Konferenz-Interessen in den folgenden Teildisziplinen: die Erste Nationen, kanadische Ureinwohner und Indigene Studien; öffentliche digitale Geisteswissenschaften; und open data Bewegung in Verbindung sind, ein. Wir heißen alle willkommen, die in den zahlreichen Disziplinen, Methodologien und Ausbildungen der Digitalen Geisteswissenschaften forschen und lehren

Folgende Typen von Einreichungen sind möglich:

- Poster (Zusammenfassung von 250 bis 500 Wörter)
- Blitzvorträge (5 Minuten. Zusammenfassung von 200 bis 250 Wörter)
- Kurzvorträge (10 Minuten, Zusammenfassung von 200 bis 500 Wörter
- Langvorträge (20 Minuten, Zusammenfassung von 750-1000 Wörter)
- Panels (90 Minuten, Zusammenfassung von 250 Wörter und eine Übersicht von 300 bis 500 Wörter)
- Forum (90 Minuten, Zusammenfassung von 500 Wörter)
- Workshops und Tutorials vor der Konferenz (2 oder 4 Stunden; 250 Wörter Übersicht und einen Unterrichtsprogramm und/oder eine Liste von relevanten bereits existierenden Tutorials)

Alle Vorschläge müssen relevante Quellen in der dementsprechenden Literatur zitieren
Die Referenzen sind nicht Teil der Wörterzählung. Bitte schauen Sie die detaillierten Anforderungen sowohl für jeden Vorschlag-Typ als auch die Vorschlag-Typen.
SIG-Gruppen werden dedizierte Treffen vor der Konferenz, die ein getrennten Gutachten-Prozess via SIG Organisatoren folgen, organisieren. Bitte kontaktieren Sie die SIG-Organisatoren bevor Sie einen Workshop oder Tutorial mit einem Thema wofür eine SIG-Gruppe existiert, einreichen.

DH2020 wird ein öffentliches blindes Begutachtungs-Verfahren benutzen. Dies bedeutet, dass Gutachter und Einreicher nicht untereinander erkennbar sind. Die Gutachten werden nur für den

Einreicher, die Gutachter und die Mitglieder des Programm-Komitees zugänglich sein.
Die Einreichungen werden nach folgenden Kriterien evaluiert:

- Gesamte Organisation und Klarheit des Beitrags (20%)
- Ausdrückliche Auseinandersetzung mit der relevanten Fachwissenschaft, mit Referenzen und Annotationen, die eine gute Kenntnis des aktuellen Wissenstands zeigen soll (30%)
- Relevanz für das Thema: Kreuzungen: amerikanische /kanadische Ureinwohner, Fachgebiete der indigenen und ersten Nationen- Fächern, öffentliche digitale Geisteswissenschaften oder open data Bewegung (10%)
- Klarer theoretischer, methodologischer oder didaktischer Rahmen und explizite Begründung (20%)
- Anwendbarkeit, Relevanz und Wert der theoretischen, methodologischen und/oder praktischen Beitrags zu digitalen Geisteswissenschaften im Allgemeinen (20%)

Die Alliance of Digital Humanities Organizations (ADHO) setzt es sich zum Ziel, eine sichere, respektvolle und kollegiale Konferenzumgebung zu schaffen, zum Nutzen aller, die an der Konferenz teilnehmen, und zur Förderung der Forschung in und der Beschäftigung mit Gebieten, die von den Organisationen, die die ADHO bilden, unterstützt werden. Von allen Autoren wird erwartet, dass sie den Verhaltenskodex respektieren. Gutachter werden gebeten die Hinweise für Gutachtern zu folgen.

STIPENDIEN FÜR NACHWUCHSWISSENSCHAFTLER_INNEN
Die Alliance of Digital Humanities Organizations bietet eine begrenzte Anzahl von Stipendien für am Anfang ihrer Laufbahn stehende Wissenschaftler_innen an, die bei der Konferenz vortragen. Die Bewerbungsmodalitäten werden zu Beginn des Jahres 2020 auf der Webseite von ADHO bekannt gegeben. Die Stipendien werden nur für Poster und Vorträge vergeben
Fragen bezüglich dieses Aufrufs senden Sie bitte direkt an die DH2020 Programm-Vorsitzenden, Laura Estill und Jennifer Guiliano gebeten beim pc2020@adho.org
–
CFP translated by Cristina Vertan with Walther v. Hahn

CFP: Italiano

Digital Humanities 2020: "Incroci/intersezioni"

L'Alliance of Digital Humanities Organizations (ADHO) invita a presentare contributi per la conferenza annuale che si terrà presso l'università di Carleton e l'Università di Ottawa presso la città di Ottawa, Canada, dal 20 al 25 luglio 2020. Si accettano proposte in algonquin, inglese, francese, tedesco, italiano e spagnolo.

CFP: algonquin, inglese, francese, spagnolo, tedesco, italiano

Presentazione dei contributi: https://www.conftool.pro/dh2020/

Scadenza presentazione dei contributi: 22 ottobre 2019 11:59 PM EDT
Notifiche di accettazione: 1 marzo 2020

Date della conferenza:
Workshop e tutorial pre-conferenza: 20-21 luglio 2020
Conferenza: 22-24 luglio 2020
Escursioni ADHO e workshop post-conferenza: 25 luglio 2020

Il tema della conferenza del 2020 è "Incroci/intersezioni". Sono particolarmente gradite proposte relative a temi di interesse locale quali gli studi indigeni e nativo americani, le public digital humanities e il movimento open data. È tuttavia possibile inviare contributi anche su altri aspetti o campi delle Digital Humanities.

Si possono presentare:

- poster (abstract di 250-500 parole);
- Interventi brevi (5 minuti; abstract di 200-250 parole);
- articoli/interventi brevi (10 minuti; abstract di 250-500 parole);
- articoli/interventi lunghi (20 minuti; abstract di 750-1000 parole);
- Sessioni pluri-contributo (90 minuti; abstract di 250 parole + 300-500 parole di descrizione);
- forum (90 minuti; abstract di 500 parole);
- workshop e tutorial pre-conferenza (2 o 4 ore; 250 parole di descrizione, un syllabus e/o tutorial esistenti rilevanti);

Durante la fase di sottomissione, si richiede agli autori e alle autrici di includere riferimenti bibliografici pertinenti. Le citazioni non vanno considerate nel conteggio delle parole. Per maggiori dettagli, si prega di consultare i requisiti per tipo e modello di proposta.
La valutazione degli incontri pre-conferenza tenuti dai gruppi SIG verrà condotta separatamente dagli organizzatori degli incontri stessi. Si prega di contattare gli organizzatori dei SIG prima di inviare una proposta di workshop e/o tutorial se l'argomento già coperto da gruppi.

In linea con il tema della conferenza, DH2020 valuterà le proposte attraverso un processo di open peer review, una formula che prevede la pubblicazione, insieme alla proposta, anche dei nomi degli autori e dei revisori. Le review restano riservate agli autori, ai revisori e ai membri del Comitato di Programma.

La valutazione delle proposte verrà condotta sulla base dei seguenti criteri:

- adeguatezza dell'organizzazione e chiarezza del testo (20%)
- riferimenti espliciti alla bibliografia pertinente con esposizione dello stato dell'arte dei relativi ambiti (30%)
- rilevanza al tema della conferenza, "Incroci/intersezioni": studi relativi ai nativi americani e ai popoli indigeni, public digital humanities o il movimento open data (10%)
- quadro teorico, metodologico e pedagogico chiaro ed esplicita dichiarazione d'intenti (20%)
- applicabilità, impatto e valore del contributo teorico, metodologico e/o pratico alle Digital Humanities (20%)

L'Alliance of Digital Humanities Organizations (ADHO) vuole creare un ambiente sicuro, rispettoso e collegiale per il benessere di tutte e tutti coloro che vi parteciperanno e per il progresso della ricerca e del sapere scientifico nei campi sostenuti dalla nostra organizzazione costituente. Tutte le persone che invieranno contributi o che parteciperanno alla conferenza devono attenersi e sostenere il codice di comportamento disponibile online. Ai revisori sarà richiesto di seguire le linee guida per i revisori.

BORSE PER GIOVANI RICERCATORI

L'Alliance of Digital Humanities Organizations offrirà un numero di borse per giovani ricercatori che parteciperanno come relatori alla conferenza. Dettagli e linee guida per presentare la domanda saranno disponibili sul sito di ADHO all'inizio del 2020:http://www.adho.org. Le borse verranno offerte soltanto per la sottomissione di poster e relazioni.

Per qualsiasi domanda relativa a questa CfP, si prega di contattare le Chair di Programma Laura Estill e Jennifer Guiliano a pc2020@adho.org.

CFP translated by Greta Franzini with collaboration by Tiziana Mancinelli, Federico Meschini and Angelo Mario del Grosso.

Conference Themes

The theme of the conference is "carrefours/intersections," a place where paths cross. The theme recalls the geographical and cultural heritage of Ottawa, a bilingual city in unceded Algonquin territory. Three specific sub-disciplinary interests will guide our use of the theme: First Nations, Native American, and Indigenous Studies; public digital humanities; and the open data movement.

First Nations, Native American, and Indigenous Studies is an interdisciplinary field exploring the history, culture, politics, issues, and contemporary experiences of indigenous peoples. As a constellation of humanities and social science disciplines, the conference should be guided by the following principles:

- be grounded in land, region, and community
- recognize the sovereign status of First Nations, Native American, and Indigenous Communities
- be polyvocal, recognizing that important community voices must be incorporated in the conference
- problematize language and hierarchical structures including being reciprocal with the community involved in the conference
- be tied to contemporary peoples and events including serving the needs of the community
- recognize indigenous data sovereignty as key to the continuing vitality of Native communities
- analyzes the historical and continued impacts of colonialism, postcolonialism, and hegemony

Public Digital Humanities reflects many of these same values:

- start with humans, not technologies or tools
- be continuously co-constructed
- encourage inclusion rather than exclusion; intellectualism rather than anti-intellectualism
- support a lack of hierarchies
- encourage kindness within critical engagement
- note the importance of working collectively, positively, and with generosity
- supports positive and fruitful partnerships with the public beyond the academy

The open data movement encourages us to:

- enable the dissemination of scholarship both in-process and in its final form
- share with one another
- assign attribution and credit fairly
- consider Open Access, Open Source, Open Scholarship, and Open Standards practices as it relates to digital projects and their research outputs
- encourage universal participation of all potential conference attendees
- consider economic accessibility of the conference

Information related to Digital First Nations, Native American, and Indigenous Studies is drawn from the Digital Native American and Indigenous Studies project as well as the University of Ottawa and Carleton University First Nations programs. Information associated with Public Digital Humanities theme is credited to Jesse Stommel, The Public Digital Humanities, Disrupting the Digital Humanities.

Types of Proposals

Proposals may be of seven types: (1) poster presentations; (2) lightning talks; (3) short presentations; (4) long paper presentations; (5) multiple-paper panels; (6) forums; (7) pre-

conference workshops and tutorials. Based on peer review and its mandate to create a balanced and varied program, the Program Committee may offer acceptance in a different category from the one initially proposed.

Conference participants may not present in more than two paper or panel sessions. If more than two submissions considered for acceptance include the name of an individual presenter or co-presenter, then during the final review phase the chair of the program committee will contact that participant and ask them which submission(s) they will withdraw to avoid exceeding that limit. This limit does not pertain to poster or lightning talk submissions. All formats are subject to peer review and should not be considered lesser than any other presentation format. Instead, the format selected should match the state of the proposed submission.

Poster Presentations

Poster proposals (abstract maximum: 250-500 words) may describe work on any specific topics or methods or present projects and software tools in any stage of development. Posters will be 24×36 inches (610×915 mm). We are unable to provide access to monitors, electrical outlets, or furniture to support laptops and other technologies. Poster presentations are intended to elicit conversations and offer opportunity to exchange ideas one-on-one with attendees. Presenters are expected to remain with their poster for the duration of their scheduled poster session. Submissions in this category are strongly encouraged.

Lightning Talks

Lightning talk proposals (abstract maximum: 200-250 words) are dedicated to a 5 minute presentation of a single project, idea, technology, or problem. It is intended to either solicit feedback from peers or to advertise the release of a new project, dataset, or tool. Submissions in this category are strongly encouraged.

Short Presentations

Short presentation proposals (abstract maximum: 250-500 words) are intended to be dynamic 10-minute presentations appropriate for reporting on works in progress, limited scholarly interventions, or for describing a singular tool or project. Short-paper submissions seek to open dialogues among scholars working on related topics. Short presentations are eligible for the Fortier Prize, which explicitly recognizes early career scholars' work.

Long Presentations

Proposals for long presentations (abstract maximum: 750-1000 words) should deal with substantial completed research, report the development of new methodologies; or present rigorous theoretical, speculative, or critical discussions. Individual presentations will be allocated 15 minutes for presentation and 5 minutes for questions. Long submissions seek substantive feedback and discussion of the submission's relationship to other scholarship in the field. Long presentations are eligible for the Fortier Prize, which explicitly recognizes early career scholars' work.

Panels

Panels (abstract maximum: 250 words for overview, plus 300-500 words for each paper) should focus on a single theme and be inherently coherent in presenting a substantial body of research

or a research question. A panel should be conceived as a 90-minute session of four to six speakers. Those submitting proposals for panels are advised to ensure that the constitution of the panel reflects the constitution of the field and/or research topic that is being addressed and ADHO's expressed commitment to diversity or to explicitly address problems in those areas.

In case the proposer's own network is too limited, the Program Committee can advise them prior to submission on whom to contact to broaden the panel. Please contact the PC chairs if you need advice.

Forum

Forums (abstract maximum: 500 words) should focus on a single thematic or methodological challenge and be designed to facilitate a conversation at large with the digital humanities community. The forum should be conceived as a 90-minute discussion with at least three facilitators. Those submitting proposals for forums are advised to ensure that the proposal explicitly addresses how attendees will contribute.
In case the proposer's own network is too limited, the Program Committee can advise them prior to submission on whom to contact to broaden the panel. Please contact the PC chairs if you need advice.

Pre-Conference Workshops and Tutorials

Tutorials and workshops are either 1) a two-hour session or 2) a four-hour session on specific techniques, software packages, or theoretical approaches with a small number of participants.

All proposals should include:

- title and brief description of the content or topic and its relevance to the digital humanities community (not more than 250 words);
- full contact information for all tutorial instructors or workshop leaders, including links to relevant cvs or teaching materials;
- description of target audience and expected number of participants (based, if possible, on past experience);
- requirements for technical support, including software installation (the conference will handle traditional technological support, but workshop organizers are expected to manage specific needs such as access to software, servers, etc.).
- a syllabus for the proposed workshop or tutorial (the syllabus may include relevant readings that may be assigned, tutorials or training guides, and/or an outline of learning outcomes and assignments).
- preferred duration: two hours or four hours

Workshops are expected to be self-financing in terms of hardware and software needs as well as coffee and lunch for participants. Participants in pre- and post-conference workshops and tutorials are required to register for the full conference as well as pay an additional fee to the conference for the workshop. Convenors of the pre- and post-conference workshops and tutorials are required to register for the conference. Please ensure that the constitution of the workshop reflects the constitution of the field and/or research topic that is being addressed; ADHO's expressed commitment to diversity; or explicitly addresses problems in those areas. In case the proposer's own network is too limited, the Program Committee can advise them before submission on whom to contact to broaden the panel. Please contact the PC chairs if you need advice.

The Planned Keynotes for DH2020

Dr. Tahu Kukutai

Dr. Kukutai specializes in Māori and Indigenous population research and leads the NIDEA research programme Te Para One E Tū Mai Nei: Māori and indigenous futures. Her research spans a broad range of population topics from iwi / tribal demography and indigenous data sovereignty, to Māori-migrant relations and the impacts of colonization on Indigenous health. Much of Kukutai's research uses mixed methods and involves working with other social scientists in Aotearoa and internationally. In addition to her work in Māori and Indigenous demography, Tahu leads the Ethnicity Counts? project which studies how governments around the world count and classify populations by ethnic-racial criteria and citizenship status.

Dr. Kukutai is a demographer specializing on Indigenous population research. She is a founding member of the Māori Data Sovereignty network. She has published extensively on Indigenous Data Sovereignty, the ethical implications of open data (particularly for vulnerable populations) and the ways in which data is used by settler colonists to enact harm.

Ry Moran

Mr. Ry Moran is the director of Canada's Centre for Truth and Reconciliation, which continues the work of the Truth and Reconciliation Commission of Canada. Dr. Moran guides the development, growth and reach of the NCTR. At the Truth and Reconciliation Commission of Canada (TRC) he served as the Director of Statement Gathering and the National Research Centre. On the TRC's behalf, he facilitated the gathering of nearly 7,000 video/audio-recorded statements of former Residential School Students and others affected by the Residential School system. He was also responsible for gathering the documentary history of the Residential School system from more than 20 government departments and nearly 100 church archives – millions of records in all.

DH2020 Keynote Selection Process

For DH2020 we received a number of individual keynote nominations and plenary panel nominations. Those names, brief biographies, links to pertinent scholarly information, and statements of nomination were augmented by a closed nomination process completed by the local organizers and program committee members. Cumulatively, then, there were four local organizers, twenty program committee members, the program committee co-chairs, and anyone who nominated through the open nomination who were solicited to provide first round of keynote and plenary panel suggestions. For more details on the keynote selection process, see dh2020.adho.org/keynote-selection-process/

Workshops

How to Read All Languages Translation Alignment with Ugarit 33
Diversity and Inclusion for Digital Humanists 33
Greening the Digital Humanities 33
Workshop on Modelling and Maintaining Research Applications in TOSCA 34
Making Research Data Public Workshopping Data Curation for Digital Humanities Projects 37
Open, Digital Scholarship Issues, Initiatives, and Research Commons in the Humanities and Social Sciences 39
Exploring the Atlas of Digitised Newspapers Enhancing Access to and Collaborative Research with Digitised Historical Newspapers 40
Sustaining Digital Humanities Centers, Labs and Institutes 42
Multimodal publication platforms for the performing arts 42
RISE and SHINE Introduction to an API-based e-infrastructure for interoperable textual resources and research tools 43
Minimal TEI Publishing with CETEIcean 43
From Sketching to Coding Teaching Visualization as a Thinking Process 44
Reimagining Graduate Education in the Digital Humanities From Below 44
Critical Project Management for DH Rethinking the Project Charter 45
Creating Digital Collections with Minimal Infrastructure Hands On with CollectionBuilder for Teaching and Exhibits 46
Culture Analytics Workshop Networks 46
TwinTalks 3 Understanding and Facilitating Collaboration in DH 47

Posters

Effect of Promotion, Rank, and Tenure Guidelines on Open Data Distribution 59

Social media in political engagements in Africa A study of Twitter use in the intra-party elections of political parties in Ghana 59

An online course system easy to make, preserve, and promote critical thinking 59

How Do Research Data Develop? International Standardisation of Scientific Data in Historical Studies 60

Improving machine translation literacy to facilitate and enhance scholarly communication 60

Faust Times Eighteen A Network Analysis of Theatre Plays Around the Myth of Faust 61

Who's Afraid of the Big Bad ? Researching Trends in the Early Gothic Novel using Word Embeddings 62

A Flow for Digitizing Japanese Historical Materials and their Long-Term Use 62

Mind the gap Filling gaps in cuneiform tablets using Machine Learning Algorithms 62

Integrating Intertextual Search into Your Web Application The Tesserae Intertext Service API 64

Visualisation of semantic shifts the case of modal markers 64

Analyzing Link Topology to Quantify the Degree of Planned Obsolesce in Online Digital Humanities Projects 67

From Metadata to linked Open Data and Wikidata Yemenite Hebrew Manuscripts and Wikidata 68

Enhanced Stand-off TEI Annotation with StoReCo A generic approach with the use of RDF. 69

Colorization of Illustrations in Charles Dickens' Novels Using Deep Learning 69

Analysis and Visualisation of Complex Familial Relationships in Greek Mythology 70

From Millions of Words to a Single Phrase Examination of the TXM Software in Hebrew 70

Xtralingua An open-source tool for extracting quantitative text profiles 71

Les Misérables & the CITE Architecture A Publication and Toolkit 71

"To be or not to be central" - On the Stability of Network Centrality Measures in Shakespeare's "Hamlet" 72

Representing Genres with distinctive Words 74

The DARIAH ELDAH consent form wizard 74

DH Teaching Activities as a Resource for Research Accessing the DH Course Registry Data via an API 75

Mythologies A Sonification of Concept Formation in Relation to the Idea of a University 76

INTEGRATING AN ANNOTATION TOOL INTO A VALIDATION PROCESS THE CASE OF READ-IT. 76

Beside and Beyond Visualising the Paradata and Metadata of Digitised Historical Newspapers with SKOS and LOD 77
HANDLE Get a Grip on MALLET 78
Time Information System, HuTime — A Visualization and Analysis Tool for Chronological Information of Humanities 79
Teaching Across the Digital Divide The Challenges of Doing Digital Humanities in the Community College Classroom 81
From data to visualisation Dante's Divine Comedy as a case study. 1 81
Visualizing Objects, Places, and Spaces A Digital Project Handbook 84
"Women's Writing in the Eighteenth Century Evaluating 'Representative' Corpora" 84
Distant Viewing Toolkit Software for the Analysis of Visual Culture 84
THOTH – Transcribing historical objects with tabulated handwriting 85
Re-interpreting the Paintings and Poems of the Lotus Sutra by a Co-occurrence Network The Religion and Life of Aristocratic during the Japanese Medieval Ages 85
Increasing Sacred Site Accessibility for Indigenous Elders through Virtualization Demonstrating Community Engagements at the Mesa Prieta Petroglyph Project 86
Statistical Traces of Censorship in Russian Political Thought 86
DH, Disciplinarity, and the Republic of Tweets 86
Better preserving the European Silk Heritage with the SILKNOW ontology. A CRM extension for modelizing the production process of silk artefacts 87
By the People at the Library of Congress 88
Analysis and Categorisation of Research Software in the Digital Humanities 88
No comment Addressing comment sections in web analysis 89
Colaboración recíproca - ¿Cómo los proyectos de humanidades digitales pueden retribuir a las instituciones culturales? 91
NEH Funding for Digital Projects 91
Extending the Utility of the HTRC Extracted Features Dataset Through Linked Data 91
Automatic Extraction of Poetry from Digitally Scanned Books 92
Building from Bones DH-WoGeM and Care in DH 94
Towards Paleographic Linked Open Data (PLOD) A general vocabulary to describe paleographic features 95
The Digital MedievalistPostgraduate Subcommittee Networking, Research, and Open Collaboration 97

The Petition Project Visualizations and Representation of the 13th-Century Communication Network of Edward I 97
"Empowered Minorities Differential Outcomes For Minorities Enjoying Kremlin Support" 97
Women and Power in the Later Roman Empire Diachronic and Geospatial Analysis Possibilities from TEI-Encoded Roman Legal Codes 98
Measuring and visualizing narrative structure in Jan Potocki's Manuscript Found in Saragossa 98
Integrating Historical Maps and Documents through Geocoding - Historical Big Data for the Japanese City of Edo 99
THE DIGITAL TURN IN AFRICA - ACADEMIC LIBRARIES AS INTERSECTIONS FOR DIGITAL HUMANITIES INITIATIVES 101
Zemsky Educational Establishments Development of the Database 101
Supervised Learning and the Indeterminacy of Disordered Genres 102
Taming the Data Web-Scraping and De-Duplicating Messy Multilingual Philosophy Corpora 102
Cartesianism as a Social Epidemic A Network Analysis Approach 103
Victorian Poverty in Print and Digital Forms Digitally Representing the Multimodal Publication History of Henry Mayhew's London Labour and the London Poor 103
Non–Linear Temporality in The Sound and the Fury --A Digital Humanities View 105
The semantics of the U.S. street map 105
Preservation and Ethical Re-Use of Crowdsourced Citizen Scholarship 105
Digital Humanities among the iSchools An Anlysis of DH Courses 106
Evaluating a Machine Learning Approach to Identifying Expressive Content at Page Level in HathiTrust 106
The Digital Analysis of Syriac Handwriting (DASH) Project Augmenting Manuscript Studies via Interactive Scriptcharts and IIIF 108
Russian Across Borders Towards Virtual Language-Centric DH Working Groups 109
Thinking Backwards, Walking Ahead Guiding Humanities Researchers through Digital Project Life Cycles 110
Assessing the Impact of an Undergraduate Digital Humanities Summer Research Program 110
Noh as Intermedia Interactive Multi-Scale, Multimedia Analyses of Full-Length Noh Theater Performances 111
The Public Approach to Re-represent Chinese Rural Memory in Cyberspace a Case Study on Gaoqian Village 113
Creating a New Semantic Model for Ancient Greco-Roman Prosopography-Toward a Contextual & Historical Description of the Prosopographical Data- 114
THE IMPORTANCE OF THE PRESERVATION OF HUMAN INVOLVEMENT OF SOUTH AFRICANS IN THE ANTARCTIC REGION AS THE ONLY COUNTRY IN AFRICA. 114

Paper Feature Extraction from Digital Images 114
Ten years recovering the memory of republican exile with citizen collaboration. The results of E-xiliad@s Project a perspective from the Digital Humanities and the Digital Public History. 115
Introducing an Automated Pipeline for a Browser-based, City-scale Mobile 4D VR Application Based on Historical Images 117
Reconsidering the Roman Workshop using machine learning to examine the processes behind inscribed texts 117
Toward Exploring Artist Information from Seal Images in Ukiyo-e Collections 118
Information extraction across textual corpora semi-automatic text-tagging workflow with Chinese local gazetteers 118
Stylo, a semantic writing tool for scientific publishing in Human Sciences 119
Les Ateliers de [sens public] de l'édition collaborative au livre ouvert 120
Visualizing Collocations in Religious Online Forums 123
"She Was Simply a Woman Who Was in Search of Variety" Supporting Complex Searches on The Willa Cather Archive with Orchid 127
SDC4Lit - a science data center for literature 128
An Improvement of Collaborative Digital Scholarly Edition with IIIF 129
Revue 2.0 Repenser la mission des revues savantes en sciences humaines et sociales 130
How Issues Guide The Life of News Stories 132
Communicating the TEI Across Linguistic and Cultural Boundaries 133
The Otletosphere an interactive map of the network around Paul Otlet 134
Because the world is multidimensional Annotations for storytelling with 3D objects 135
Critical Making and Public Facing DH in the Liberal Arts Classroom a case study. 136
TextAnnotator A web-based annotation suite for texts 137
The Historic Graves crowdsourcing project 10 years transcribing Irish history 139
Free Land for Whom? Mapping Land Acquisition and Dispossession under the Homestead Act, 1863-1912 141
En France, on hurle – au Québec, on braille! Une comparaison diachronique de l'usage des verba dicendidans la littérature franco-québécoise et française standard 141
Entry and Rise In Bureaucracy Patterns of Career Advancement in Joseon Dynasty 143
Using Bayesian Network Classifiers to Trace the Constuction of Race, Gender, and Class in all Print Editions of the Encyclopedia Britannica 1768-2010 143

Extracting a Social Network of Musicologists 143
Construction of the corpus of senmyō one of the oldest materials of Japanese language 146
Acquisition and Analysis of a Meme Corpus to Investigate Web Culture 147
Building Bridges Pedagogical Strategies for Introducing Digital Humanities in the Undergraduate and Graduate Classroom 150
Not Coding Alone Teaching the Collaborative Habits of DH in a MOOC 150
A Priority Programme on Computational Literary Studies 151
Tracking the digital dead A progress report in the use of transfer learning, image analysis, and forensic anthropology to classify human remains sold on social media 152
The Hierarchy of Characters Quantitative Dominance Relations in Russian Drama 153
An interactive multimedia companion to Wagner's Lohengrin encoding and visualising a motivic study 153
Computer Vision Approaches to Body Language and Facial Expressions Analysis in XX-century Cinema 155
A simple method of extracting keywords from texts 156
THE LUSOPHONE DIGITAL HUMANITIES AND WHAT THEY (WE) ARE DOING FROM THE SOUTH TEXTUAL CORPUS ANALYSIS AND FAIR PRINCIPLES TO TACKLE HEGEMONY. 156
Experiments in Analysis of Industry—Academia Collaboration and Research Trends 157
Le Lotus and the women's voices of Egypt 158
A Survey of Data and Encodings in Word Clouds 158
Network Analysis Finds Shifts in the History of Modern Architecture 158
An Open Data Approach to Revealing Indigenous Texts in Large-Scale Digital Repositories A Case-Study of Locating Pages of Māori Text in the HathiTrust 159
Making Ukiyo-e Easier to Discover A Recommender System for Digital Archives 160
Conceptual Modeling of Similarities and Duplication in Large-Scale Digital Libraries 160
Feminist Revisions An Emergent Model for DH in Libraries 161
Automatic Labeled Data Generation for Person Named Entity Disambiguation on the Ming Shilu 161
MAZI means together An open-source "minimal computing" local network infrastructure used for cultural event organisation, fieldwork research and community-based curation 162
NewsEye A digital investigator for historical newspapers 164
Text Analysis of 19th and 20th Century Etiquette Books Using R 166

A comparative study of sentiment and topics in migration related tweets *166*
Studying Geographical Patterns Across John Milton's Genres *167*
RavenSpace Books Unbound *167*
"Webs of Violence in Beowulf" *168*

Lightning Talks

Opening up Open Data Strategies & success stories 180
PRESERVATION OF OSUN-OSOGBO CULTURAL HERITAGE IN NIGERIA A DIGITIZATION DISCOURSE 180
The Rabbinic Social Network 181
Interactive Spatial Positioning of Historical Photographs in Virtual Reality Results of a User Study 183
University-Industry Partnerships in the Humanities View from the partner and academic perspective 183
Computation and Rhetorical Invention Finding Things To Say With word2vec 184
Coding Literary Ecologies 185
The Montias Case an experiment with data reconciliation and provenance between research and cultural heritage institutions 185
"Digitizing Rochester's Religions Piloting a Community–University Partnership in the Digital Humanities." 187
La Enciclopedia de Banrepcultural Una enciclopedia digital de patrimonio cultural colombiano, que nace de la colaboración de los visitantes de las Bibliotecas del Banco de la República 188
«La planète numérique» d'un peuple autochtone transnational Une analyse des liens hypertextes des sites web amazighs 188
Quantitative Intertextuality Analyzing the Markers of Information Reuse 188
DIRECT - Digital Intertextual Resonances in Early Chinese Texts 189
Bringing Newcomers into the Fold Faculty Development through Values-Driven DH Pedagogy 189
Análisis del coro como personaje en la dramaturgia grecolatina y española incluidas en DraCor 190
What Is Formalisation? And Why Do We Need to Talk About It In DH? 190
Building Online Communities as a Platform for Collaborative Learning and Cross-Cultural Self- Expression; Saudi Female Students' BlogHajj Behind the Scenes as an Example 191
Annotating spatial entities in Romanian Novels 191
Stop Words 192
Not TEI by Design A Digital Scholarly Edition of Paradise Lost 193
Haikupedia 193
Tiempo de destrucción A stylometric examination of the posthumous publication of Luis Martín-Santos's unfinished work 194
Promoting Library Accessibility and Outreach with Special Collections in Virtual Reality 194

Linking digitalized Buddhist scripts and Buddhist studies publications 194
Creating Structured and Reusable Data for Tourism and Commerce Images of Edo Using IIIF Curation Platform to Extract Information from Historical Materials. 195
Deep Learning based Attribute Representation in Ancient Vase Paintings 196
Reproducible Humanities Research Developing Extensible Databases for Recording "Messy" Categorisation, Annotation and Provenance Data 197
Crossroads of Literature and History; Text Mining the Modern Persian Literature 199
Intersecting Digital Archives and GIS Technologies Bringing IIIF Images to GIS Environments 199
Enriching Metadata for Irish Traditional Music at the American Folklife Center 199
The Fight for National Language Rights in the USSR 200
Performer communities within a hyperconnected company network 200
Creating a Meaningful Genre Schema and Metadata using IMDb data for a Large-Scale Digital Humanities Project in Media Studies 201
DIGITAL HUMANITIES SCHOLARSHIP IN AFRICA Prospects and Challenges 202
Comparative analysis of rhythms in Japanese folk songs 203
Encoding Belief why digital has become the new location of salvation and destruction. 203
Revenge of the Silk. Elaborating and using a CIDOC CRM extension a case study on silk-related artefacts 203
CLARIN for DH Scholars 204
Computational Analysis of Emotions and Topics in Survivor Interviews about Nanking Massacre 206
Does Culture Still Matter? A Cross-Cultural Perspective on Human-Robot Interaction and Artificial Intelligence 207
INVESTIGATING THE USE OF SOCIAL MEDIA FOR HEALTH CARE ADVOCACY IN NIGERIA A DIGITAL MEDICAL HUMANITIES APPROACH 208
Abusing the Concept of Normalization for Better Collation Results (and Profit) 208
Prosopography in the digital era? Experiments in managing data from a Historical and Biographical dictionary 209
The visual digital humanities – topics, researchers and cultures 209
Transatlantic Cultures a digital humanities project for a connected history of the Atlantic world (18th-21st centuries) 209
Digital Approaches to Name Disambiguation of Chinese Historical Figures 210
Diffusion and Change of the Romantic Concept in Historical Newspapers 210

Rhetorical interpretation of stylometry methods in disciplinary writing a question of hermeneutics 211
Cyborg Public Humanities The Digital-Material Exhibit and the Classroom 211
Women Personal Archives the digitization of CPDOC's feminine collection and the relevance of women political action in contemporary Brazil 211
Pricing Opinions and Tastes The Market for « Primitive » Artefacts through Sentiment Analysis 212
What Is A Reader? How Readers on Goodreadsare Changing the Canon in the Twenty-First Century 213
The Future of Storytelling in the Age of AI and Posthumanism 213
Portugueses de Papel Exploring Lusophobia and desidentificação in Brazillian literature 214
Contrasting Playwrights through Sonic Signatures 215
That's Not Vintage, It's Obsolete Steve Reich's early e-sketches as a case study of current issues with early historical music notation software and why it matters 215
Between Phallus and Freedom An Ethnography on the Embodied Experiences of Tinder Users in Cape Town 216
Sober Hopes Designing an Introductory Course for Computational Research in Hebrew / Arabic Texts 216
Sentiment sensitivity. Using sentiment analysis in literary studies to analyze genre and the depiction of illness 216
Data-Driven In-class Activities for Chinese History Pedagogy 219
Tracking the framing of politicians and news events across the multilingual Wikipedia translation and its unseen impact 219
A Prototype for a Truly Decentralized, Queryable LOD Store 220
Deep Learning for Recognizing Chinese Historical Documents 220
Co-occurrence of modal markers a network analysis approach 222
Making Uganda's Intellectual History Digital Knowledge Preservation and Ethical Considerations 222
Critical Making, Gender, and Textual Analysis Beginning a Major Project 222
Eruptive Art Digital Analyses of Atmospheric Anomalies in American Painting 223
Subject Analysis of Cultural Heritage Consuls in Taiwan and Diplomatic Incidents 223
Differential treatment how testimony is discussed in assault and sexual assault cases at the Court of Appeal of Québec 223
Leveraging Google Sheets and GitHub for Data Curation on the Princeton Ethiopian Miracles of Mary Project 224
Evaluating Semantic Relation Alignment in Historical Bibles usingHistorical Embeddings 225
An Online Odyssey Digital Storytelling in the Humanities Classroom 226

Behind the Scenes of the Spectators Uncovering Anonymous Authors in Periodicals of the Enlightenment 227
The dynamics of Czechoslovakia's policy as a reaction to the influence of the USSR and the Warsaw Pact countries on the materials of the archive of special services and the periodical press of Czechoslovakia 1968-1989. 229
Book printing and the rise of the vernacular in Europe, 1500–1820 229
Corpus-DB a Scriptable Textual Corpus Database for Cultural Analytics 230
"Network Analysis and Spatial Stylometry in American Drama Studies" (NASSA) 230
The Dutch Veterans Memoires Work Bench Bridging the gap between a traditional and computational reading of memoirs 231
TexTiles Visualizing the Patterns of Discourse 231
Victimisation and otherisation in Cameroon Anglophone literature A Study of Selected Authors from the 1980s to Present 232
Creating New Music with Big Data and Evolutionary Algorithms 233
Poetry Machines Empowering Creative Writers to Design DH Tools 234
Championing Open Interactions with RERUM; Focusing and Preserving an Accessible Scholarly Conversation Across Disciplines 235
Investigation of reliability in public history documents from 18th century 235
Explore Plato's Cave in Virtual Reality 236
"What Our Favorites Can Tell Us A Computer-Directed Reading, Topic Modeling and Trend Analysis Investigation into 2,000 New York TimesBestsellers and Genre Fiction Winners from 2005 to 2016" 236
Typicality in the U.S. Novel 236
Digital Humanities Start-Up Grant Retrospective Towards a Sustainability Index for Grant Funded DH Projects 237
The Shakespeare and Company Project 237
Computational access to library's digital collections 238
MONA from public art to our art with a mobile app 239
Enhancing Community through Open DH Website Design 241
Structures of Thought Common Readers and Data Modeling Annotations 242
Argumentation Mining in Spoken Georgian Political Language 242

Short Presentations

Booth Tarkington, Blindness, Dictation, and the Durability of Style 255
Testing Rolling.Classify 255
Ethics and responsibilities of open access - lessons learned from the Wikipedia project of the Atikamekw First Nation 255
CR/10 Website as a Digital Public Humanities (DPH) Site 256
Features of Timelessness Intermediate Report on a Quest for Stylistic Features that Mark Literary Canonicity 257
Gmail's Smart Compose A Critical Composit(ion) 258
Transing DH Adopting a Trans-Centric Approach to Building the Transgender Media Portal 259
Theoretical intersections cognitive poetics, cultural evolution, and distant reading in literary studies 259
Inscriptions of Israel/Palestine 261
The GeoNewsMiner An interactive spatial humanities tool to visualize geographical references in historical newspapers 263
Web 2.0 for an Authoritative Web Site 264
NER on Ancient Greek texts with minimal annotation 264
THE READING EXPERIENCES ONTOLOGY (REO) REUSING AND EXTENDING CIDOC CRM 265
From Public Humanities to Social Remembering Big Data and the Digital Redlining of Women in Country Music Culture 265
Project Twitter Literature Scraping, Analyzing, and Archiving Twitter Data in Literary Research 267
"Beyond the logic of commensurability a cultural analysis of media artworks and digital media in information capitalism" 268
The Plant Humanities Workbench Using Linked Open Data to Discover Early Modern Plant History 269
Bitter Wind Adapting Greek Tragedy for Spatial Computing 269
DUCAT Passage/Translation Alignment with the CITE Architecture 269
MaDiH (####) Mapping Digital Cultural Heritage in Jordan 270
Modelling Consultation Workflows for Research Data Management in the Humanities 271
Identifying relations between characters in Afrikaans, Tshiven?a, and Xitsonga book 273
Never Alone, Never Finished Defining Never Alone's World Games Genre as Ethical Alternative to Empathy Games 273

Intersections of the Cultural Kind Public Digital Humanities and the Museum 274
Using computer vision to identify graphic elements in Picasso's poetry 274
Evolución y escenario actual de las Humanidades Digitales en España 276
Datasets of Criminal Faces Within and Under Facial Recognition Software (FRS) From a Digital Humanities Perspective 279
Potential Archives How Digital Humanities and Feminist Ethical Praxis Will Transform the Interdisciplinary Artist Archive 281
Analyse automatique pour une étude du genre quels jugements des écrivaines au XIXe siècle ? 282
An Interactive 3D Visualization of RDF-based Digital Editions 282
Metadata model for an archeological data lake 283
Haiku Author Recognition 284
"Open GLAM" Opening up digital cultural heritage collections for the digital humanities 286
Towards Modeling the European Novel. Introducing ELTeC for Multilingual and Pluricultural Distant Reading 286
AI4AV (Artificial Intelligence for Audiovisual) Design and Evaluation of a Shared System for LAMs 288
Developing the Data Foundry the National Library of Scotland's data-delivery platform 289
The "No-Interface" Interface for Research VR 290
Mapping Meaning learnings from indigenous mapping technology for Australia's digital humanities mapping infrastructure 290
Unsupervised Rumor Detection on Twitter using Topic Homogeneity 291
From Silo to Repo Enforcing File Structure to Improve Workflow and Access 293
Who Teaches When We Teach DH? Some Answers and More Questions. 294
Constructing and Deconstructing Digital Serpentary Goals Established, Challenges Encountered, Lessons Learned, Discoveries Made. 295
Quantifying the Unknown How Many Manuscripts of Sévigné Still Exist? 297
Widening access to Linked Ancient World Data 301
The Miiyupimatisiiun Research Data Archives Project Co-establishing a usable Indigenous data management system. 302
Early Modern Digital Itineraries Modeling European Place and Space, 1545-1747 303
Visceral Histories Analyzing Transmission through Katherine Dunham's Dance Repertory 304

Virtual Space, Voice, and Gender in Recorded Popular Music from 2008–2018 305
Digital History of Post-Soviet Virtual Museums 305
Residual or Essential On the Problem of Overlap as Argument to Broaden the Solution Space of Markup 307
A Neural OCR Engine for North Saami 307
Visual Authority, Privacy, and Surveillance in the Age of Deepfakes 307
The Words Themselves A Content-Based Approach to Quote Attribution 308
A Conceptual Framework for the Analysis of Multilayer Networks in the Humanities 308
Analyzing Border Crossings in Contemporary Irish Literature with Literary Mapping 310
Graph models for the genesis of Goethe's Faust 310
Taking a Visual Turn in DH 313
Algorithmically Mapping Historical Fascination in Late Imperial Chinese Literature 313
Indigenous Collaborations on Digital Public Memory Work in the Classroom 314
Annotating Reader Absorption 314
OuPoCo, the combinatorial poetry workbench (L'ouvroir de poésie combinatoire) 317
An Ontology for Critical Editions of Variant Text 320
Sustainably modeling a 3D virtual tour Can we avoid the "black box"? 320
Genocidio de los pueblos originarios y modos de distribución de las tierras fiscales en la Patagonia Austral la prensa local y los regímenes de propiedad agraria (1929-1962). 321
Training Algorithms to Read Complex Collections Handwriting Classification for Improved HTR Models 321
Queering the Tape Recorder Transforming Surveillance Technologies through bill bissett's Queer Poetic Voice 322
Finding Identical Ukiyo-e Prints across Databases in Japanese, English and Dutch 324
Pulses Decentralized Smooth Semantization of Atomic Historical Knowledge 325
Encoding Working Lives Linking Labor, Office, and Religion in 18th-century Manuscript Collections 325
WoPoss - A World of Possibilities. An open access annotated corpus to study the diachrony of modality in the Latin language 327
Reconstructing the gendered division of labor in the French textile trades. Distant reading of primary qualitative sources with NLP tools (18th century-beginning of the 20th century) 327
Crowdsourced Text Transcription on the Zooniverse Platform Models for Design, Development and Evaluation 328

Teaching at the Intersection of Digital Humanities and Visualization 328
Fast Search with Poor OCR 330
Mapping the Gay Guides Understanding Queer Spaces in Pre- and Post-Stonewall America 333
Accessibility and Reception Vector Semantics, Reading Publics, and the Changing Reception of Literary Works 333
Archival Justice in the DH Classroom Examining El Diario de la Gente and the Chicanx Movement in Boulder, Colorado 333
Mapping the Ghadr Directory Dimensions of Diasporic Anticolonialism and State Surveillance 334
Crossroads, shortcuts, detours, bypasses or dead ends? Attempts to standardization and interoperability within the context of the CLARIAH-DE project 334
DIALOGUES ON DETENTION - AN ITERATIVE, INCLUSIVE APPROACH TO 3D VISUALISATIONS OF COLONIAL CAMPS IN KENYA 335
Exploring Guangxu-era missionary activities in Taiwan from Chinese Recorder, Dan-Hsin Archives and Ming-Qing Taiwan Administrative Archives through DocuSky 336
Semi-Automatic Identification of Travelogues 336
Modeling possible stories. A digital edition of Giovanni Domenico Tiepolo's Divertimento per li Regazzi 339
Doing Ecological Identity Ecoparadox in Governor Babatunde Fashola's Speeches 341
Data Art, Public Engagement, and Luso-Brazilian Literature and Culture Reflections on the Bird-watching Exhibition 341
"On the record" transcribing and valorizing qualitative interviews with XML-TEI 342
Crowdsourcing Training Data Efficacy and Ethics 344
Foundations of Distant Reading. Historical Roots, Conceptual Development and Theoretical Assumptions around Computational Approaches to Literary Texts 344
New Scholarship in the Digital Age Making, Publishing, Maintaining, and Preserving Non-Traditional Scholarly Objects 346
Counting Burned Books A Quantitative Analysis of the Great Fire of London 346
Reading Like Alice Tricks of Perspective in Reading and Visualization 347
De-centering Darwin opening up nineteenth century scientific correspondence and putting Darwin in context 347
Digital Linguistic Landscape the multilingual system of Ottawa 347
Digital Public History Origins of the Field and Possibilities for the Future 348
A Collaborative Workspace for Archival Research MIA and the EURONEWS Project 348
Data-driven analysis of canonical works in early modern Britain 350

Seven collections of prosody the difficulty of visualizing non-hierarchical data 350
Aural Palimpsests Voices, Space, History 352
Reutilización digital de datos públicos de la BNE como método heurístico para el análisis, mapeo y evaluación de las políticas culturales artísticas locales; aplicación a los sistemas expositivos temporales de Málaga y Valencia. 353
(Re)collecting the Archive Virtually Recovering Medieval Manuscripts Destroyed During World War II 357
How Bisexual+ Community Members Tweet about Health A Community-Informed, Topic Modeling Approach to Characterizing Under-Reported Health Needs of LGBTQ Twitter Users 358
Inventing A Digital Chorapleth Map An Alternative Choragraphic/Psychogeographic Mapping Praxis 358
Physically-Based Rendering (Maya) for Experimental Modeling of Medieval Objects 359
Sounding Spirit Overlays and Understories OCR and Consortial Thematic Research Collections 359
Mapping Topic Evolution Across Poetic Traditions 359
"Quantifying Kissinger" - Visualizing the Many Worlds of the Kissinger Correspondence 363
Harry Potter and the Engaged Reader Community Interactions and Influence in Serialized Fan Fiction 363
A tale of two web archives Challenges of engaging web archival infrastructures for research 364
A digital assessment tool for monitoring and planning food security interventions in rural households of Uganda 365
Machine Learning for Literary Criticism Analyzing Forms, Genres, and Figurative Language 365
A TEI-XML Model for tone representation in African languages 366
Beyond Translation language hacking and new pathways into language. 367
How Do You Spell "??"? The expansion of Unicode and the blurred line between text and image in digital space 367
La obra de los científicos sociales del Siglo XX llevadas a las plataformas digitales del siglo XXI los principios y prácticas de las humanidades digitales 368
CollectionBuilder Using Static Web Technologies to Create Digital Collections and Teach DH Principles 368
Stopifu Supporting Task-Specific Stoplisting for Topic Models 369
Mapping Memory in Post-Hurricane Maria Puerto Rico By Combining Rhetoric and GIS 369
Interactive Transcripts and Multimodal Shakespeare 369
Metrics for Distant Reading Literary History 370
Visualizing a Translational Queer Poetics 371
Caveat magister a computational approach to designing a Latin curriculum 371

Human-Centered Computing for Humanists Case Studies from the Computational Thinking and Learning Initiative at Vanderbilt University 372
Close-Reading the Endnotes for Word-Frequency Signposts & Polyvocal Citations 374
The Potential and Danger of Digital History Pedagogy 374
Cherokee Syllabary Texts Digital Philology and Experiential Learning 374
Inferring book relationships at the trillion-word scale 375
Can Digital Humanities Help in Finding Research Questions? A Comparative Analysis of the Attitudes Towards Neo-Confucianism Study of the Scholars Today and 300 Years Ago 375
Discovering Digital Humanities Methods Through Pedagogy A Study of Introductory DH Syllabi 375
Constructing and Analyzing Short Science Fiction at Scale 376
Archives, Podcasts, and Place-Based Audio Storytelling 377
A Digital Study of Ralph Ellison's Integrative Form 378
Digital History, Deep Mapping and Public Heritage in the Undergraduate Classroom 378
Implementation of Event Classification and Spatio-temporal Information Integration Analysis based on Meteorological Records in Historical Texts A Digital Humanities Case Study on Climate Change and Social Trends 380
An Advising Framework for Intersecting Disciplines within Graduate Digital Humanities Certificate Programs 381
A Visualization-Assisted Reading Systemfor a Neo-Confucian Canon 381
Circulating and Computing Anti-Colonialism Modeling Revolutionary Information Networks and Radical Ideas in the Long 1960s 383
Linked Archives a Digital Approach to Visualising Historical Literary Networks 383
DH and the Gig Economy The opportunities and ethical challenges of Sourcery 384
Designing accessible platforms Site-specific digital literature for people with visual impairment 384

Long Presentations

What is a Genre? A Graph Unified Model of Categories, Texts, and Features 394
Restoration and Repurposing of DH Legacy Projects 395
How can science and knowledge be created for all and by all without #linguisticjustice? Findings from a two-year study on the intersections between citizen science, social media, crowdsourcing, and Translation Studies. 397
Algorithms of Resistance Using OCR and AI for Social Justice 397
The place of models and modelling in Digital Humanities Intersections with a Research Software Engineering perspective 398
The Open Knowledge Program Creating Space for Digital, Public Scholarship 401
What's in a Face? Gender representation of faces in Time, 1940s-1990s 404
Metaphors we read by Finding metaphorical conceptualizations of reading in web 2.0 book reviews 405
Prototyping the SF Nexus Collaborative Models for Digitizing and Curating Speculative Fiction Collections as Data 409
Sociocultural trend signatures in minimal persistence and past novelty 410
Mind the Gap Bridging Distant and Close Reading across Heterogeneous Text Collections 411
"Exalting the Cult of Gentlemanly Amateurism" [1] Improving Computational Analysis with Humanities Narrative-Building Methodologies 411
Trans-Hispanic Networks of Feminist Solidarity The Rise and Spread of #8M 414
Tribesourcing Southwest Films Counter-Narrations and Speaking Back to Colonial Rhetoric 414
Situated Materialities, Object Agency and Resistant Bodies in Subalternized Craft Neoliberal Market/Tourist Logics and Questions around what/why of the "decolonializing" of digital humanities 416
Locating Biographies of Canadian and Native American Women Nationality, Race, and Mobility in North America 416
Framing the experience a study of the history of interfaces to digital humanities projects 417
The Semantics of Structure in Large Historical Corpora 417
Understanding Depth in Deep Learning Knowledgeable, Layered, Impenetrable 421
Opening the Stage -- A Quantitative Look at Stage Directions in German Drama 422
StarCoder A general neural ensemble technique to support traditional scholarship, illustrated with a study of the post-Atlantic slave trade 425
When Classical Chinese Meets Machine Learning Explaining the Relative Performances of Word and Sentence Segmentation Tasks 425

"Telling bigger stories" Formal ontological modelling of scholarly argumentation 428
DH Awards Possibilities and problems of an openly-nominated and openly-voted DH awareness-raising activity 431
Distant, Deep, Thick Reading across Languages Building towards Inclusive Digital Epistemologies of Text 432
Variorum Editions Across Text and Music the Early Modern Songscapes Project 432
PoetryLab. An Open Source Toolkit for the Analysis of Spanish Poetry Corpora 433
Linking Time, Space, and Statements in One GIS System A Use Case of Studying Individuals' Biographies 436
Towards Open Computer Vision Methods Visual Analytics of Machine Learning Models in the Digital Humanities 439
AdA Filmontology – a machine-readable Film Analysis Vocabulary for Video Annotation 443
Topics, buckets, and psychiatry. On the collective creation of a corpus exploration tool 445
Illuminating Past Labor Making Transformation Processes of Historical Documents Visible 447
Using smart annotations to map the geography of newspapers 452
Historical Newspaper Content Mining Revisiting the impresso Project's Challenges in Text and Image Processing, Design and Historical Scholarship 454
Serving the city an automatic information extraction for mapping Amsterdam nightlife (1820-1940) 455
All papers are data papers from open principles to digital methods 457
Data Visualizations of Non-Human Animals and their Land 459
The Explainability Turn Critical Digital Humanities and Explanation 459
Binding Media. Contemporary Electronic Literature and Experimental Publishing 461
Contextualizing Victorian Newspapers 461
Conceptual Analysis as Thickening Influence as an Example 466
Multilayer Network Applications in Poetry. Centralities & Automated Corpus Expansion in Computationally Assembled Anthologies 468
Quantifying cultural change literary translation in Mexican periodicals (1894-1931) 468
Null models in authorship analysis - an alternative approach to established methods in stylometry 470
Legal Literacies for Text Data Mining 470
What if we can't manage our way to better sustainability? 472
Reassembling Elephants A Multi-Spatiotemporal Visualization Method for History and Humanities Data 472
Reading Certainty Evidence from a Large Study on NLP and Witness Testimony 475

Novel Worldbuilding Science Fiction 477
Teaching Digital Humanities on Raspberry Pis. A Minimal Computing Approach to Digital Pedagogy. 480
Recognizing Indigenous Sovereignty and Imagined Futures in the Context of Linked Data 482
"Work In Progress" Public Humanities, Digital Humanities, and Forms of Labor, Collaboration, and Audience 482
Gamergate Predicting the Present 482
Scrollership a new name for some new ways of writing 482
Morgan Papers Exploring the Correspondence of California's First Female Architect 483
Speech Rhythm and Syntax in Poetry and Prose 483
Dynamic Systems for Humanities Audio Collections The Theory and Rationale of Swallow 487
The Digital Curation of Broadcasting Archives at the Canadian Broadcasting Corporation Curation Culture and Evaluative Practice 491
Nanohistory – An Experimental Digital History Methodology 494

Panels

Understanding Rhyme Through Network Analysis 500
Linked Art Networking Digital Collections and Scholarship 504
Scribes, Scholars, & Scripts creating a Digital Humanities community through crowdsourcing 509
Replication and Computational Literary Studies 509
Constructing Spatial Narratives Considerations and Practices Across Communities 514
Global Renderings in the Queer Digital Humanities 515
Digital Humanities Methods for Analyzing Holocaust and Genocide Testimonies 515
Compartir lo que nos une. Digitizing and Curating Colonial Records from the Caribbean and Central and South America for Public Outreach 516
Risking I.T. Trans-Feminist Queer Digital Performance, Archives and Information Technologies 520
Working at the Intersection Digital Humanities Pedagogy, Social Justice, and Librarianship 521
Twitter is an Indigenous Territory 524
Cultural Analytics and the Book Review Models, Methods, and Corpora 525
Applied DH Lessons from Using the Methods and Technology of Digital Humanities to Build Tools and Services for Researchers, Librarians and Curators 526
Sustainability in the Making and Knowing Project From Scholarly Edition to Community Crossroads 527
Sustainability Strategies for Digital Humanities Systems 530
#whereislatinxdh 537
The Web in Amber How the Ephemeral Human Digital Experience is Preserved and Analyzed 537
Deciphering Encoded Music Increasing Access to Open Data Resources 538
Unmaking/Remaking Memory Work Centering Community Narratives of Latinx Lived Experience 539
Operationalizing Surveillance Studies in the Digital Humanities 541
Markup as Critical Pedagogy 542
Ethical Soundings in Collaborative Digital Humanities Research Projects Critical Scenarios from The SpokenWeb 542
Towards an Indian Decolonial Digital Humanities 548
Designing and Gaming Doing Digital Humanities in India 551

Forums

Exploring the Undiscovered Contours of DH 558
Who needs tool directories? A forum on sustaining discovery portals large and small 558
Sharing is Daring Global Copyright Laws, Knowledge Protection, and the Impact on Open Digital Humanities 559
Quantitative Text Analysis Intersections with Feminism, Postcolonialism, Critical Race Theory and Queer Theory 560
Ensuring Software Quality and Sustainability of DH Projects 561
Towards a Multilingual DH 561
Using and Sharing Crowdsourced Data 562
Sunny Days Sweepin' the Clouds Away Critiquing the Digital Technologies Operationalized in Smart Cities 562
The Open Lab Digital Scholarship and the Question of Community 563
Representation Matters How Do We Promote a Radically Inclusive Digital Humanities? 563
Asian American Digital Humanities Building Radical Communities through Justice-Oriented Praxis 564
Building new global DH communities Africa and beyond 564
FAIR Data Principles for the Humanities and Social Sciences Sharing Knowledge and Perspectives Between Europe and Canada 565
Exploring Conferences as Sites of Public Digital Humanities in Times of Crisis 565
Demystifying ADHO, the Alliance of Digital Humanities Organizations 566
Should We Burn it Down? The State of Librarians in Digital Humanities Pedagogy, From Uriah Heeps to Unicorns 567
Pedagogy of the Digitally Oppressed Practicing Anti-colonial DH Pedagogy and Research 567

Workshops

How to Read All Languages: Translation Alignment with Ugarit

Chiara Palladino
chiara.palladino@furman.edu
Furman University, United States of America

Anna Zhang
Furman University, United States of America

Maryam Foradi
Digital Humanities, University of Leipzig, Germany

Tariq Yousef
tariq@informatik.uni-leipzig.de
NLP Group, University of Leipzig, Germany

Keywords: translation alignment, digital pedagogy, language learning
Topics: Comparative (2 or more geographical areas), English, BCE-4th Century, 5th-14th Century, 15th-17th Century, electronic literature production and analysis, public humanities collaborations and methods, Literary studies, Translation studies

This workshop will illustrate the importance of translation alignment in the field of slow reading and language learning. We will provide a short theoretical overview on the principles of translation alignment, together with a hands-on tutorial on Ugarit (http://ugarit.ialigner.com/), a web-based translation alignment editor. Ugarit is designed as a Citizen Scienc
e
tool, aiming at collecting training datasets of manually aligned words from diverse text corpora. The ultimate goal of Ugarit is to improve automatic translation alignment methods and to implement a set of dynamic lexica, with particular regard for languages with less supported infrastructures. However, the tool also has a strong pedagogical potential, which has been assessed in the course of various hands-on workshops and in an ongoing integration in school curricula: we have tested how translation alignment with Ugarit can help readers to engage with languages that they have never seen, grasping their essential semantic and morphological aspects. We propose text alignment as a way to empower the perception of the complexity of a language, but also as a method to leverage usual obstacles in the process of reading apparently "impenetrable" sources by directly engaging with them. At present, Ugarit includes aligned pairs from 36 languages (including less represented languages, like Bulgarian, Ethiopic, Sanskrit, Yiddish, and Armenian), 277 unique users, and about 23,400 parallel texts hosted.

Diversity and Inclusion for Digital Humanists

Barbara Bordalejo
barbara.bordalejo@usask.ca
University of Saskatchewan, Canada

Daniel Paul O'Donnell
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

Keywords: Diversity, Inclusion, Bias, Privilege
Topics: Global, English, Contemporary, digital activism and advocacy, Disability and differently-abled studies, Feminist studies

This workshop highlights issues of diversity, inclusivity, and collaboration in Digital Humanities.

Through practical exercises and dialogue, we build a safe atmosphere for the discussion of strategies to isolate obstacles preventing diversity and offer solutions for the development of inclusive environments.

As part of our work, we developed The Privilege Game, used to emphasize and showcase the many different kinds of privilege derived from our society's power structures and to create awareness among practitioners of the contrasts to be found in the rich and ever-growing space of the Digital Humanities.

We cover topics such as "implicit bias," "cultural cloning," privilege, intersectionality and solutions to be implemented in the creation of safe and inclusive environments.

Greening the Digital Humanities

Kit Heintzman
kheintzman@fas.harvard.edu

Harvard University, Canada

Keywords: carbon emissions, environmentalism
Topics: Global, English, Contemporary, eco-criticism and environmental analysis, sustainable procedures, systems, and methods, History of science

This workshop would introduce audiences to the following issues: recent historiography attention to the internet's carbon footprint, a case study in a semester long system of tracking the carbon emissions of a digital humanities course, trnasferable pedagogical strategies with regard to eco-conscious digital scholarshp, and critical reflections upon issues of "individualism" in discourse of digital consumption patterns versus system changes.

Workshop on Modelling and Maintaining Research Applications in TOSCA

Philip Schildkamp
philip.schildkamp@uni-koeln.de
Data Center for the Humanities, University of Cologne, Germany

Lukas Harzenetter
lukas.harzenetter@iaas.uni-stuttgart.de
Institute of Architecture of Application Systems, University of Stuttgart, Germany

Frank Leymann
Institute of Architecture of Application Systems, University of Stuttgart, Germany

Brigitte Mathiak
Data Center for the Humanities, University of Cologne, Germany

Claes Neuefeind
Cologne Center for eHumanities, University of Cologne, Germany

Uwe Breitenbücher
University of Stuttgart, Germany

Keywords: Living Systems, Research Applications, Software Stacks, Sustainability, TOSCA
Topics: Global, English, Contemporary, software development, systems, analysis and methods, sustainable procedures, systems, and methods, Humanities computing, Informatics

Abstract

The project "SustainLife - Sustaining Living Digital Systems in the Humanities" that is currently running at the Institute of Architecture of Application Systems (IAAS, University of Stuttgart) and the Data Center for the Humanities (DCH, University of Cologne) deals with the conservation of research applications in the field of Digital Humanities (DH). By employing the TOSCA standard (Topology and Orchestration Specification for Cloud Applications) to fully automate the deployment of DH applications and to keep them available in the long term, we try to tackle the problem of software obsolescence in the field of DH. To interactively demonstrate our approach to the international DH community, we would like to give a workshop on the topic "Modelling and Maintaining Research Applications in TOSCA" in the run-up to the DH 2020 conference. Thereinwe will show how to model (DH) software systems with TOSCA and share experiences and best practices on how to work with the OpenTOSCA ecosystem, an open-source implementation of the TOSCA standard.

The Problem

The establishment of the DH as an independent scientific research area as well as the increasing usage of digital methods in the research process require adjustments to common result assurance practices. For example, the long-term archiving (LTA) of primary research data uses well-established practices such as employing standardized data formats and forwarding data to permanent repositories. However, the fact that digital artifacts generated in DH-oriented research do not only consist of primary data but also contain research software is mostly disregarded (Sahle and Kronenwett, 2013). Moreover, the variety of DH research outcomes includes so-called "living systems" in which the software to present, access or analyze the data represents an essential part of the actual research output (Bingert et al., 2016). In contrast to classical research results such as monographs or encyclopedias, living systems cannot be served long-term without maintenance as their instantiation, supervision, and permanent provisioning represent major technical, organizational, and financial challenges. Furthermore, the heterogeneity of the research software generated in the DH requires a highly flexible preservation strategy,

i.e., a suitable technology that ensures standardization, reusability, and archiving of as many digital artifacts as possible (Barzen et al., 2018). In addition to the aforementioned challenges, i.e., heterogeneity, underfunding, and obsolescence of digital artifacts, scientific practice requires long-term interoperability and traceability of all research outcomes. With regard to digital systems, these requirements are (1) constant accessibility, (2) the possibility of error-free operation, and (3) the ability to reconstruct any stage of development of a research application at any time without major structural difficulties.

Our Approach

The TOSCA standard (OASIS, 2013 and 2019) allows software systems to be modelled, provisioned, and deployed in a standardized and provider-independent manner. Thus, it is suitable for long-term archiving and operation of research applications produced within the field of DH (Neuefeind et al., 2018 and 2019). Following the TOSCA standard, applications are modelled in "Topology Templates" by describing their components and their relations amongst each other: Components are represented as "Node Templates", while relations are modeled as "Relationship Templates". Moreover, the semantics of a Node Template or Relationship Template are dictated by reusable types, i.e., "Node Types" and "Relationship Types" respectively. For example, a Python web application can be modelled as a Node Template that is an instance of the "Python Application" Node Type. To express that the Python Application accesses a MySQL database, a second Node Template that is of type "MySQL Database" can be added to the Topology Template. Then, the connection between both components can be described by a Relationship Template that is an instance of the Relationship Type "connectsTo". Additionally, to specify that both components are running on an Ubuntu virtual machine (VM), a Node Template of type "Ubuntu VM" can be added, while Relationship Templates of type "hostedOn" between the Python Application Node Template and the VM Node Template, as well as between the MySQL Database and the VM describe their respective hosting relations.

Hereby, TOSCA's type system enables the modelling of reusable component types, e.g., the "Python Application" Node Type, which can be reused in multiple Topology Templates describing different applications. Therefore, synergic effects emerge as existing Node Types can be reused in other Topology Templates, easing the modelling of new applications. In addition, the open-source TOSCA implementation OpenTOSCA (Breitenbücher et al., 2016) offers the possibility to graphically model applications using the TOSCA editor "Winery" (Kopp et al., 2013) which further simplifies the creation of new applications by providing drag-and-drop modeling capabilities.

Workshop Curriculum

During our four hour workshop, we will (1) give an overview to different solutions for long-term preservation of living systems and (2) describe the modeling language TOSCA. Based on these theoretical units, practical tasks will introduce (3) the modelling of an existing application using TOSCA and (4) how applications can be deployed using the OpenTOSCA ecosystem. Thus, by combining the theoretical foundations and the practical application of TOSCA, the participants will be able to model (research) software systems according to the standard and provision and deploy applications using the OpenTOSCA ecosystem.

The practical tasks are structured as follows: (1) Identify the components of an application and (2) describe them and their relations among each other in an TOSCA-based application topology, i.e., in a Topology Template. By fragmenting an application into its components and mapping them to TOSCA Node Types, the Topology Template describing the application can then be modelled using the OpenTOSCA ecosystem. Afterwards (3), the modelled TOSCA application will be deployed by the OpenTOSCA runtime. Moreover, by sharing our experiences and best practices in using OpenTOSCA with the community, we will introduce concepts such as "software stacks" in a practical way.

Target Group

The workshop is primarily designed for data center employees, libraries and other institutions focusing on infrastructures for long-term archiving and operation of heterogeneous software systems. Previous experience in dealing with Linux and writing shell scripts as well as with software stacks and service orchestration are helpful but not necessary for a successful participation. To provide a productive context for communicating the described content and to enable individual consultation and support, we designed the workshop for about 20 participants but limit it to a maximum of 30 participants.

Technical Prerequisites

For a successful participation in the workshop, it is necessary that each participant brings his/her own laptop. Although a shared instance of the OpenTOSCA ecosystem will be provided, it is desirable that all participants set up an OpenTOSCA instance on their work equipment prior to the workshop in order to perform modelling and deployment tasks on their own devices. Therefore, registered participants will be provided with all necessary information about system requirements and how to setup OpenTOSCA prior to the workshop. Furthermore, relevant documentation, publications, and manuals will be provided both in

advance and in the context of the workshop. In addition, a stable internet connection as well as a sufficient number of power outlets for all electronic devices are indispensable.

About the Instructors

Uwe Breitenbücher is a research staff member and postdoc at the Institute of Architecture of Application Systems (IAAS) at the University of Stuttgart, Germany. His research vision is to improve cloud application provisioning and application management by automating the application of management patterns. Uwe was part of the CloudCycle project, in which the OpenTOSCA Ecosystem was developed. His current research interests include cyber-physical systems, blockchains, and microservices.

Anna Fischer is a research assistant at the Data Center for the Humanities (DCH) at the University of Cologne and joined the "SustainLife" Project in January 2020. Her recent research and working activities have focused on data management and software development for natural language processing tasks, e.g., in collaboration with one of the chairs for Romance linguistics at the University of Cologne.

Lukas Harzenetter is a research associate at the Institute of Architecture of Application Systems (IAAS) at the University of Stuttgart, Germany. He received his Master of Science degree from the University of Stuttgart in Software Engineering in 2018. His research interests are in the field of cloud deployment and management models focusing on the development and change of such models over time. Lukas is part of the "SustainLife" project which is working on sustainable application deployments in the domain of digital humanities.

Frank Leymann is a full professor of computer science and director of the Institute of Architecture of Application Systems (IAAS) at the University of Stuttgart, Germany. His research interests include service-oriented architectures and associated middleware, workflow- and business process management, cloud computing and associated systems management aspects, and patterns. Frank is co-author of more than 400 peer-reviewed papers, about 70 patents, and several industry standards. He is an elected member of the Academy of Europe.

Brigitte Mathiak is chairwoman of the Data Center for the Humanities (DCH) and is particularly interested in data management and text mining. The idea for the "SustainLife" project arose after she had experienced again and again how living systems have to be abandoned or neglected. She is Junior Professor for Digital Humanities at the University of Cologne and Senior Scientist at the Leibniz Institute for the Social Sciences (GESIS).

Claes Neuefeind is a postdoc at the Cologne Center for eHumanities (CCeH) at the University of Cologne. He worked with Philip Schildkamp and Lukas Harzenetter on the DFG-LIS project "SustainLife" until October 2019 and changed for a position that is responsible for coordinating the Digital Humanities of the North Rhine-Westphalian Academy of Sciences and the Arts office.

Philip Schildkamp has been researching since 2015 and teaching since 2017 at the University of Cologne. He studied sociology, psychology, and Digital Humanities information processing. The main topics of his employment are technical infrastructure measures in the field of (Digital) Humanities and the orchestration of distributed software systems. Since March 2018, Philip has been part of the DFG-LIS project "SustainLife" at the Data Center for the Humanities (DCH).

Acknowledgements

This poster is partially funded by the DFG-LIS project "SustainLife" (GEPRIS 379522012).

References

Barzen, J. and Blumtritt, J. and Breitenbücher, and JKronenwett, S. and Leymann, F. and Mathiak, B. and Neuefeind, C. (2018). SustainLife - Erhalt lebender, digitaler Systeme für die Geisteswissenschaften. In: Book of Abstracts of the 5th annual Conference of the Digital Humanities im deutschsprachigen Raum (DHd 2018), pp. 471-474.

Bingert, S. and Blumtritt, J. and Buddenbohm, S. and Engelhardt, C. and Kronenwett, S. and Kurzawe, D. (2016). Anwendungskonservierung und die Nachhaltigkeit von Forschungsanwendungen. In: Forschungsdaten? ?in? ? den? ?Geisteswissenschaften? ?(FORGE? ?2016),? pp. 14-16.

Breitenbücher, U. and Endres, C. and Képes, K. and Kopp, O. and Leymann, F. and Wagner, S. and Wettinger, J. and Zimmermann, M. (2016). The OpenTOSCA Ecosystem. Concept & Tools. In: European Space Project on Smart Systems, Big Data, Future Internet. Towards Serving the Grand Societal Challenges. Volume #1, pp. 112-130.

Kopp, O. and Binz, T. and Breitenbücher, U. and Leymann, F. (2013). Winery - A Modeling Tool for TOSCA-based Cloud Applications. In: Proceedings of the 11th International Conference on Service-Oriented Computing (ICSOC 2013), pp. 700-704.

Neuefeind, C. and Harzenetter, L. and Schildkamp, P. and Breitenbücher, U. and Mathiak, B. and Barzen, J. and Leymann, F. (2018). The SustainLife Project - Living Systems in Digital Humanities. In: Proceedings of the 12th Advanced Summer School on Service-Oriented Computing (SummerSoC 2018) (IBM Research Report RC25681), pp. 101-112.

Neuefeind, C. and Schildkamp, P. and Mathiak, B. and Mar?i?, A. and Hentschel, F. and Harzenetter, L. and Breitenbücher, U. and Barzen, J. and Leymann, F. (2019). Sustaining the Musical Competitions Database. A TOSCA-based Approach to Application Preservation in the Digital Humanities. In: Book of Abstracts of the 29th Digital Humanities Conference (DH 2019), https://dev.clariah.nl/files/dh2019/boa/0574.html (retrieved: 2019-09-10).

OASIS (2013). Topology and Orchestration Specification for Cloud Applications Version 1.0, http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html (retrieved: 2019-09-10).

OASIS (2019). TOSCA Simple Profile in YAML Version 1.2, http://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.2/TOSCA-Simple-Profile-YAML-v1.2.html (retrieved: 2019-09-10).

Sahle, P. and Kronenwett, S. (2013). Jenseits der Daten. Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner Data Center for the Humanities. In: LIBREAS. Library Ideas #23, pp. 76-96.

Making Research Data Public Workshopping Data Curation for Digital Humanities Projects

Felicity Tayler
ftayler@uottawa.ca
University of Ottawa, Canada

Sarah Simpkin
sarah.simpkin@uottawa.ca
University of Ottawa, Canada

Marjorie Mitchell
Marjorie.Mitchell@ubc.ca
University of British Columbia Okanagan

Constance Crompton
constance.crompton@uottawa.ca
University of Ottawa, Canada

Karis Shearer
karis.shearer@ubc.ca
University of British Columbia Okanagan

Matthew Lincoln
mlincoln@andrew.cmu.edu
Carnegie Mellon University Libraries

Mikhel Proulx
mikhel.proulx@concordia.ca
Concordia University

Meghan Goodchild
meghan.goodchild@queensu.ca
Queen's University and Scholars Portal, Ontario Council of University Libraries

Keywords: data communities, research data management, data curation, digital scholarship, best practices
Topics: English, North America, Contemporary, data publishing projects, systems, and methods, digital ecologies and digital communities creation management and analysis, Library & information science

A lack of formal training opportunities for data curation in multi-site DH teams means that the data produced in these teams is in danger of being lost! This four-hour workshop will cover all areas of data management including: IP permissions and informed consent, data collection, metadata standards, file sharing, preservation (data deposit), and data sharing through the open data spectrum of access. Participants will work on their own data curation challenges in break-out sessions and with reference to case study examples presented by a panel of DH scholars and digital asset management specialists: Constance Crompton (uOttawa), Karis Shearer (UBCO), Matthew Lincoln (Carnegie-Mellon U), Mikhel Proulx (Concordia U and Indigenous Digital Art Archive), Meghan Goodchild (Queen's U and Scholars Portal). The lesson plan is designed and delivered by Felicity Tayler (uOttawa), Sarah Simpkin (uOttawa), and Marjorie Mitchell (UBCO).

This workshop is a good preparation for researchers who must create a data management plan to comply with funding agency requirements. The workshop arises at a moment when DH researchers have greater access to funding to support large-scale multi-partner projects with diverse digital assets. The manifold nature of DH, and its reflexive challenges to culturally imposed power imbalances in digital systems presents unique challenges for data curation. Responding to

the conference thematic of cultural and disciplinary intersections, this workshop proposes that DH is one of the social and conceptual spaces where the informal networks of international "data communities" arise through acts of data curation and sharing (Cooper and Springer). Our approach to data curation recognizes that data communities are multilingual and multi-cultural just as they cross epistemological and disciplinary lines.

Contact information of co-facilitators

Felicity Tayler, Research Data Management Librarian, University of Ottawa ftayler@uottawa.ca | https://uniweb.uottawa.ca/members/3883

Example of workshop previously led on this topic: "Data Curation for Communities of Sound: A Speculative Workshop" led by Felicity Tayler and Marjorie Mitchell. SpokenWeb Symposium. Simon Fraser University, BC. 30 May 2019; OLA SuperConference 2020, Toronto, ON .30 January 2020. https://www.olasuperconference.ca / conf_workshop/data-curation-for-communities-of-sound-a-speculative-workshop/

Sarah Simpkin, Head, Research Support (Arts and Special Collections), University of Ottawa sarah.simpkin@uottawa.ca https://orcid.org/0000-0003-0774-4777

Marjorie Mitchell, Copyright, Scholarly Communications, and Research Data Management Librarian, Research Services, University of British Columbia Okanagan Marjorie.Mitchell@ubc.ca | https:// spokenweb.ca/

Contact information of confirmed presenters/ affiliated authors

Constance Crompton, Canada Research Chair in Digital Humanities in the Department of Communication, University of Ottawa ccrompto@uottawa.ca | http://constancecrompton.com

Karis Shearer, Associate Professor, English and Cultural Studies, University of British Columbia Okanagan karis.shearer@ubc.ca | http://amplab.ok.ubc.ca/

Matthew Lincoln, Research Software Engineer, Carnegie Mellon University Libraries mlincoln@andrew.cmu.edu | https://matthewlincoln.net/

Mikhel Proulx, Indigenous Digital Art Archive, Concordia University, mikhel.proulx@concordia.ca | http://indigenousfutures.net/archive/ Meghan Goodchild, Acting Head, Discovery and Technology Services, Research Data Management Systems Librarian, Queen's University and Scholars Portal, Ontario Council of University Libraries (OCUL), meghan.goodchild@queensu.ca | https://www.actorproject.org/

Description of target audience

This workshop is of interest to DH researchers working in national and international collaborative teams, in particular, those who are funded through agencies that have Data Management Plan requirements. Students and Research Assistants will also benefit from this training, as the next generation of researchers who will adopt these data curation best practices. As a key focus of the conference is First Nations, Native American, and Indigenous Studies (FNAIS), we anticipate that many workshop attendees will bring knowledge systems into the room with them that challenge data curation best practices based in Western traditions of property ownership and colonial methods of data collection - we will address these gaps.

Technical requirements

Should this workshop be accepted, the University of Ottawa Library has committed to hosting us in its "Tinkering Lab," designed for collaborative learning styles and data visualization. The room is equipped with an 8K LED video wall, and 2 LED 4k displays, as well as capacity for up to 50 participants seated in modular group learning dynamics. We will provide our own technical support. Refreshments will be served in the mezzanine area outside of the Lab.

Syllabus for proposed workshop

Making Research Data Public: Workshopping Data Curation for Digital Humanities Projects Prepared by: Felicity Tayler, Sarah Simpkin, Marjorie Mitchell

Learning Outcome:

By the end of this session, participants will be able to understand data curation terms and best practices, in a way that is meaningful to their own disciplinary and computational methodologies.

Learning Objectives:

Participants will be able to 1) Recognize and describe their research data 2) Identify data curation practices they already undertake 3) Describe data management best practices that they could adopt to better achieve their goals 4) Summarize what they have learned as it applies to their own projects

Lesson Content / Process:

40 mins

What is your data? Introduction to data curation key concepts (Presenters: Felicity Tayler, Sarah Simpkin, Marjorie Mitchell)

Introduction to data management plans in international contexts (DMPOnline (UK), DMPTool (USA); DMP Assistant (Canada)

Activity: Ask group about what data they are producing/working with.

---

60 mins with Q&A

"Learning from examples" DH data curation successes (and failures) presentation of 1-4 case studies

Constance Crompton: LINCS: Linked Infrastructure for Networked Cultural Scholarship

Karis Shearer: Press Play: Making Spoken Web Research Data Public

Matthew Lincoln: Humanities Data: The Middle Path for Real-Life Researchers

Mikhel Proulx: Indigenous Data Sovereignty and Ethical Allyship in the Archive

Meghan Goodchild: Collaborating on ACTOR (Analysis, Creation, and Teaching of Orchestration)

---

30 min

Refreshments & networking

---

30 min

"What is Your Data Flow and Discovery Model?" How to map out your research data at different phases of the project addressing: IP permissions and informed consent, data collection and storage, metadata description, file sharing, preservation (data deposit), data sharing through the spectrum of access, publication, credit and citation.

Activity: Facilitator-led break-out sessions (F. Tayler, S. Simpkin, and M. Mitchell). Participants work individually or in a group.

---

10 min

Break

---

30 min

"What is Your Spectrum of Data Access?"

How to map your data flow models onto 5 categories of access from secure & protected to open license.

Data Papers & Data Journals

Activity: Facilitator-led break-out sessions (F. Tayler, S. Simpkin and M. Mitchell). Participants work individually or in a group.

---

10 min

Break

---

30 min

Sharing & feed-back session

Participants synthesize what they have learned by presenting their data flows in small groups for feed-back

Some relevant readings and resources

Anderson, Katrina and et al. "Student Labour and Training in Digital Humanities." DHQ: Digital Humanities Quarterly 10.1 (2016): n.p. Web. .

Institute for Critical Indigenous Studies. "Episode 2: Indigeneity in DH." Recoding Relations. Web.

Canadian Writing Research Collaboratory. LINCS Cyberinfrastructure Project. 2019. Web.

Carnegie Mellon University. The Digital Humanities Literacy Guidebook. 2019. Web. https://cmu-lib.github.io/dhlg/

Cooper, Danielle and Rebecca Springer. "Data Communities: A New Model for Supporting STEM Data Sharing." ITHKA S+R (2019): 1-26. Web. .

Data Curation Network. Data Curation Network Primers. 2018. Web. .

Digital Curation Centre. About DMP Online. 2019. Web. .

Maryland Institute for Technology in the Humanities. Digital Humanities Data Curation Guide. [2014]. Web.

McGovern, Nancy Y. "Radical Collaboration and Research Data Management: An Introduction." Research Library Issues 296 (2018): 6-22. Web. .

Posner, Miriam. "What's Next: The Radical, Unrealized Potential of Digital Humanities." Gold, Matthew K. and Lauren F. Klein. Debates in the Digital Humanities 2016. Minneapolis: University of Minnesota Press, 2016. 32-41. Print.

Portage Network. DMP Assistant. Web.

Tayler, Felicity; Marjorie Mitchell, and Rebecca Springer. "Emergent Data Community Spotlight II: An interview with Felicity Tayler and Marjorie Mitchell on the Spoken Web Project." Ithaka S+R. Web. 10 September 2019. < https://sr.ithaka.org/blog/emergent-data-community-spotlight-ii/>

University of California Curation Centre. DMP Tool : Data Management General Guidance. Web. 2019. .

Wernimont, Jacqueline and Elizabeth Losh, eds. Bodies of Information: Intersectional Feminism and the Digital Humanities. Minneapolis: University of Minnesota Press, 2018.

Open, Digital Scholarship Issues, Initiatives, and Research Commons in the Humanities and Social Sciences

Alyssa Arbuckle
alyssaa@uvic.ca
University of Victoria, Canada

Ray Siemens

siemens@uvic.ca
University of Victoria, Canada

Keywords: open scholarship, open access, research commons, digital scholarship
Topics: English, North America, Contemporary, open access methods, public humanities collaborations and methods, Communication studies

In the spirit of the public humanities, initiatives are emerging that foster the open sharing, re-purposing, and development of scholarly projects, publications, educational resources, data, and tools. One example is the Canadian Humanities and Social Sciences (HSS) Commons, an in-development prototype for a national-scale, online research commons that features:
subject repository for open access publications with digital object identifiers (DOIs) upon upload and FAIR (Findable, Accessible, Interactive, Reusable) guidelines for data managementproject development environment integrated with Google Drive, Github, or Dropboxindividual user profiles with federated login/identity authorization, including ORCIDblogging capabilitiessubject interest groups and member interactions (e.g., profile building, messaging)

The workshop includes:
talks by key players in the Canadian digital research infrastructure world;a theoretical overview of the commons in research development and community building;instruction in how to set up a profile, upload publications, form a special interest group, and begin a digital scholarship project in the prototype Canadian HSS Commons.

This initiative builds on consultations among Implementing New Knowledge Environments (INKE) Partnership members-in particular the Canadian Social Knowledge Institute, Compute Canada, and the Federation for the HSS-and has roots in similar initiatives such as Humanities Commons, a partner in this work.

This platform is intended to serve the unique needs of the Canadian HSS community. Based on early-stage consultations we anticipate that it will draw a significant number of DH practitioners. This prototype is built on HUBZero-an open source content management system available for deployment within other international contexts as well.

Exploring the Atlas of Digitised Newspapers: Enhancing Access to and Collaborative Research with Digitised Historical Newspapers

Melodee Beals
m.h.beals@lboro.ac.uk
Loughborough University, United Kingdom

Emily Bell
e.bell@lboro.ac.uk
Loughborough University, United Kingdom

Keywords: digitisation, newspapers, metadata, history, literature

Topics: Global, English, 19th Century, 20th Century, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), metadata standards, systems, and methods, History, Library & information science

Building upon the two-year Digging into Data project, *Oceanic Exchanges: Tracing Global Informatio n Networks in Historical Newspaper Repositories, 1840-1914* (http://www.oceanicexchanges.org), this workshop will introduce participants to the Atlas of Digitised Newspapers, a comprehensive guide to the histories, structures and metadata of the digitised newspapers collections studied by the project, including those held by: Chronicling America,
The Hemeroteca Nacional, Digital de México,
The British Library, The Times Digital Archive, Delpher, Europeana, Suomen Kansalliskirjaston
Digitoidut Sanomalehdet, Trove, Papers Past, ZEFYS

The *Atlas* provides readers with a deep contextualisation of these collections as well as detailed technical information about how to obtain, interpret, manipulate and map metadata and content across collections. Discussions will be supported by two additional online resources-an interactive visualisation of the metadata mapping and a linked dataset to support cross-database research.

Starting with humans rather than technologies or tools, the *Atlas* specifically uncovers the people

behind the metadata and selection decisions. The workshop demonstrates a commitment to continuously co-constructing this work, and we hope to foster future collaborations. The Atlas will encourage research that "analyzes the historical and continued impacts of colonialism, postcolonialism, and hegemony"; siloed national archives privilege collections that look inward, but *Oceanic Exchanges* has used the combined collections to explore multilingual/multicultural groups with future possibilities for work on comparative Indigenous studies.

The aims of the workshop are to:
Present the findings and outputs of the project; Create a space for critical discussion of the findings and future research opportunities by participants of the workshop; Allow for initial collaborations and hands-on use of the outputs in developing historical, literary, and computer science research questions.

All participants will have the opportunity to submit and review the following in advance of the workshop: An online profile of previous work with digitised newspapers and historical periodicals.Suggestions for possible in-workshop collaborations using the Atlas, visualisations and LOD datasets across our core expected disciplines.

Learning Outcomes

By the end of the workshop, participants will:
Develop a nuanced understanding of available data beyond OCR text;Practice cross-collection data management and cleaning;Develop an understanding of metadata and selection practices.

Schedule

Introductions of workshop organisers and participants, Introduction to Oceanic Exchanges,
Presentation of the Atlas of Digitised Newspapers, Discussion of collections not
represented in the Atlas by participants, Walkthrough of Datasets and Use Cases, Introduction of Disciplinary Challenges, Division into groups for Disciplinary Challenges, Presentations, feedback and discussions, Closing Remarks

Suggested Readings:

Newspaper Preservation

Bourke, Thomas A. "Scholarly Micropublishing, Preservation Microfilming, and the National Preservation Effort in the Last Two Decades of the Twentieth Century: History and Prognosis." Microform Review 19.1 (1990): 4-16. https://doi.org/10.1515/mfir.1990.19.1.4.

Silverman, Randy. "Retaining hardcopy papers still important in digital age." Newspaper Research Journal 36.3 (2005): 363-72. https://doi.org/10.1177/0739532915600749.

Walravens, Hartmut, ed. International Newspaper Librarianship for the 21st Century. Müchen: K. G. Saur, 2006.

Newspaper Digitisation

Abruzzi, Ray, Luisa Calè, and Ana Parejo Vadillo. "Gale Digital Collections: Ray Abruzzi Interviewed by Luisa Calè and Ana Parejo Vadillo." 19: Interdisciplinary Studies in the Long Nineteenth Century 21 (2015): http://19.bbk.ac.uk/articles/753/.

Sigauke, D. T. 'Digitisation Technologies for Newspaper Archives in Zimbabwe: The ICT Requirements for Digitising a Selected Bulawayo Newspaper Publication at the National Archives of Zimbabwe'. In 2017 IST-Africa Week Conference (IST-Africa) (2017): 1-10. https://doi.org/10.23919/ISTAFRICA.2017.8102315.

King, Edmund. "Digitisation of Newspapers at the British Library." The Serials Librarian: From the Printed Page to the Digital Age 49.1-2 (2005): 165-81. https://doi.org/10.1300/J123v49n01_07.

Metadata Standards

Europeana Data Model. Europeana Pro. Accessed 11 October 2019. https://pro.europeana.eu/resources/standardization-tools/edm-documentation.

Wilkinson, M. D. et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific Data 3 (2016): article number 160018. https://www.nature.com/articles/sdata201618

Crowdsourcing Heritage Data

Recker, Mimi M. and David A. Wiley. "A Non-authoritative Educational Metadata Ontology for Filtering and Recommending Learning Objects," Interactive Learning Environments 9.3 (2001): 255-71. https://doi.org/10.1076/ilee.9.3.255.3568.

Workshop leaders:

M. H. Beals (m.h.beals@lboro.ac.uk, http://mhbeals.com/CurriculumVitae/) is a Lecturer in Digital History at Loughborough University specialising in the intersection of media and migration in the eighteenth and nineteenth centuries. She is the leader of the Ontologies work package of Oceanic Exchanges and PI of the "Empowering Users of Historical Digitised Newspapers Collections' project at Loughborough University.

Emily Bell (ebell@lboro.ac.uk, https://lboro.academia.edu/EmilyBell) is a Research Associate in Digital Humanities at Loughborough University on the 'Empowering Users of Historical Digitised Newspapers Collections' project, which aims to enable libraries to improve the choice and availability of historical newspapers, supporting a more informed technical and ethical use of collections by the public.

Program Committee:

More information on the full Oceanic Exchanges team can be found at https://oceanicexchanges.org/team/

Sustaining Digital Humanities Centers, Labs and Institutes

Katherine L. Walter
kwalter1@unl.edu
University of Nebraska-Lincoln

Paul Arthur
paul.arthur@ecu.edu.au
Edith Cowan University

Carrie Heitman
heitman@unl.edu
University of Nebraska-Lincoln

Keywords: Centers, Labs, Institutes, Sustainability
Topics: Global, English, Contemporary, digital ecologies and digital communities creation management and analysis, sustainable procedures, systems, and methods, History, Library & information science

In this two hour tutorial, the facilitators will raise questions that are germane to digital humanities centers, labs and institutes regardless of location, size, or stage of development. These begin with the general question of how can digital humanities support institutional goals, such as cross-department collaboration, diversity and interdisciplinary research, while at the same time being distinctive and forward looking. The intention is to facilitate discussion about these key issues and others, such as the cultural context in which centers operate, and the opportunities of working within different national research systems and what can be learned from these. Our goal is to encourage participants to share their very different experiences.

Among discussion points will be funding and sustainability, which kinds of funding arrangements are possible, and what models exist to inform those starting a new center, lab or institute or retooling an existing center, lab or institute. This will lead to a consideration of the difference between infrastructure and operations funding vs. funding for research or pedagogy. Other discussion points include the importance of building a digital humanities community, ideas for ensuring wide engagement, the value of creating policies and a governance structure, and developing partnerships and connections through professional associations and diverse communities. We will also spend some time discussing how centers can be a catalyst for developing skills in graduates for new careers, international exchanges, visiting scholarships and postdoctoral research fellowships.

Multimodal publication platforms for the performing arts

Clarisse Bardiot
clarisse_bardiot@mac.com
Valenciennes University, France

Miguel Escobar Varela
m.escobar@nus.edu.sg
National University of Singapore, Singapore

Keywords: Multimedia annotation, sustainability, performing arts
Topics: Asia, Global, Europe, English, Contemporary, annotation structures, systems, and methods, sustainable procedures, systems, and methods, Performance Studies: Dance, Theatre

Writing about and studying the performing arts, especially for the contemporary period, means using performance video-recordings. However, these recordings are rarely communicated with the intended audience, except as supplementary materials in academic journals and books. But would it be possible to develop scholarly interfaces that interlink text and video in more complex ways? The challenges are institutional rather than purely technical. We need to agree on standards for annotation, publication, review and preservation. But every project in the performing arts is unique. Conceptual discussions in the performing arts have often been described as a "sophisticated disagreement" (Strine, Long, and Hopkins 1990). Would it be possible to extend this approach, from the conceptual to the technical? Can we develop a technical and institutional framework for sophisticated disagreement that enables the

multimodal annotation annotation of performances and integrates this into a publication pipeline?

In this workshop, we explore different tools for annotating and communication audiovisual, multimedia scholarship. We conduct a survey of existing projects and give an in-depth introduction to MemoRekall (Bardiot) and a recent framework for integrating interlinked text and video in ways that are easy to preserve and that have wide applicability across domains, which is used as the basis of the Asian Intercultural Digital Archiva (Escobar Varela). Both are open-source and free applications.

The workshop is open to people with an interest in performing arts scholarship and to everyone looking for ways to develop multimodal publication platforms.

RISE and SHINE: Introduction to an API-based e-infrastructure for interoperable textual resources and research tools

Sean Wang
swang@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Keywords: API, research infrastructure, research software, digital research tool
Topics: Asia, Comparative (2 or more geographical areas), Global, English, Contemporary, digital research infrastructures development and analysis, software development, systems, analysis and methods, Humanities computing, Library & information science

Textual resources still overwhelmingly exist in digital silos with limited interoperability amongst themselves and to digital research tools. Some large-scale infrastructural projects are intended to address this issue by centralizing resources, while others focus on promoting interoperable standards for exchange of textual resources. In this four-hour workshop, we will present our modular and decentralized solution to these issues: RISE and SHINE. We will have hands-on sessions on how to make use of our technical products with two primary target groups: research tool developers who would like to learn how to make RISE-linked texts available in their tools, as well as resource owners who would like to learn how to make their texts discoverable in a SHINE-compatible format.

Minimal TEI Publishing with CETEIcean

Raffaele Viglianti
rviglian@umd.edu
University of Maryland, United States of America

Hugh Cayless
hugh.cayless@duke.edu
Duke University, United States of America

Keywords: text encoding, publishing, static sites, minimal computing
Topics: Europe, English, North America, Contemporary, digital publishing projects, systems, and methods, physical & minimal computing, Design studies

This half-day workshop will introduce CETEIcean (pronounced /s##ti:#n/), a JavaScript library for displaying Text Encoding Initiative XML in a web browser. CETEIcean's key feature is the isomorphic transformation of TEI into HTML by registering TEI elements with the browser as Custom Elements. This is in contrast to the typical methods for presenting TEI on the web that use the XML stack (XSLT/XQuery) to convert TEI XML sources to simple HTML-in the process discarding much of the information in the sources. Loading a TEI document with CETEIcean makes all of its information content accessible to the browser: TEI Custom Elements can be styled and made interactive with CSS and JavaScript just like HTML. We will begin with a discussion to contextualize this approach within the "minimal computing" movement, particularly as pioneered by the GO:DH Minimal Computing Working Group. This approach fosters "reconnecting with our knowledge production" in order to think critically about the question "what do we need?" (Alex Gil 2015 https://go-dh.github.io/mincomp/thoughts/2015/05/21/user-vs-learner/). Identifying the minimal technical stack for running a TEI project can help content creators to reach their goals quickly and effectively, even when there are constraints on the resources they can access or the computing environments they rely upon. Minimal computing principles argue that this can be an essential instrument

in giving voice to underfunded and underrepresented groups. Attendees will work on a simple site publication embedding either a provided TEI example or their own TEI document.

From Sketching to Coding: Teaching Visualization as a Thinking Process

Eric Alexander
ealexander@carleton.edu
Carleton College, United States of America

Adam Bradley
adam.bradley@uoit.ca
Ontario Tech University, Canada

Catherine DeRose
catherine.derose@yale.edu
Yale University, United States of America

Mennatallah El-Assady
mennatallah.el-assady@uni-konstanz.de
University of Konstanz, Germany Ontario Tech University, Canada

Uta Hinrichs
uh3@st-andrews.ac.uk
University of St Andrews, United Kingdom

Stefan Jänicke
stjaenicke@imada.sdu.dk
University of Southern Denmark, Denmark

Keywords: critical visualization, visualization literacy, pedagogy
Topics: Global, Europe, English, North America, Contemporary, curricular and pedagogical development and analysis, spatial & spatio-temporal analysis, modeling and visualization, Computer science, Education/ pedagogy

Discussions of data visualization or visual analytics often focus on the underlying software and tools. When it comes to teaching others to create and interpret visualizations, however, we need to stress the theory and decision-making processes that feed into the resulting visual representations. This workshop provides strategies for teaching visualization literacy, from sketching by hand to creating visualizations using computational tools. Led by instructors from the humanities, visual analytics, design, and computer science, sessions will cover different methodologies and challenges of teaching visualization to audiences with varied technical and disciplinary expertise and goals.

The workshop walks participants through hands-on exercises, reflecting on how tools (both analog and digital) can enhance but also constrain our thinking. We will present visualization as a thinking process, a way to engage critically with data rather than only as an end product. In an increasingly visual and data-driven society, we must teach visualization as a discourse in order for it to be applied and interpreted responsibly. We will highlight critique as a pedagogical technique, as well as a method for spotting and mitigating miscommunication and misinterpretation. Encompassing interactive introductions to visualization theory, visualization creation, and design activities, this workshop will provide opportunities to discuss topics related to engaging with visualization within and beyond academic settings. Participants will leave the workshop with resources, ideas, and sample exercises they might use in their own practice and teaching.

Reimagining Graduate Education in the Digital Humanities From Below

Simon Appleford
simonappleford@creighton.edu
Creighton University, United States

Gabriel Hankins
ghankin@g.clemson.edu
Clemson University, United States

Anouk Lang
anouk.lang@ed.ac.uk
University of Edinburgh, United Kingdom

Keywords: graduate education, postgraduate education, graduates, early career researchers, alt-ac

Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, curricular and pedagogical development and analysis, Education/ pedagogy

This workshop welcomes all participants with an interest in graduate education and training in data-oriented work in the humanities. It seeks to provide a space in which current models from a range of national and institutional contexts can be shared, evaluated, critiqued and reimagined, with a particular emphasis on those who have experience of graduate education 'from below': students, practitioners and scholars from visible minorities and other groups who are not well represented in positions of institutional power where decisions about graduate education are made. In addition to those with teaching and program leadership responsibilities, we are particularly keen to include current graduate students, early career scholars, those who are or have been precariously employed, those in alt-ac positions, and professionals working outside the academy in fields where digital humanities graduates might be hired.

The workshop is intended not as a teacher-led tutorial, but rather as an intensive conversation in which lightning talks will be interspersed with writing sprints, such that engaging discussions can be quickly sparked and insights captured through periods of focused writing.

Please note that a grant proposal is currently under consideration by the Royal Society of Edinburgh which, if successful, will enable us to provide bursaries to cover the workshop fee for ten participants, and provide a contribution of around $CAD650 towards the travel costs of four of those participants. These bursaries are intended to help those from countries outside the US and Canada to attend the workshop who might otherwise find the costs prohibitive.

Critical Project Management for DH Rethinking the Project Charter

Natalia Ermolaev
nataliae@princeton.edu
Princeton University, United States of America

Rebecca Munson
rmunson@Princeton.EDU
Princeton University, United States of America

Andrew Janco
ajanco@haverford.edu
Haverford College, United States of America

Keywords: Projects Process Ethics Collaboration
Topics: Global, English, Contemporary, project design, organization, management, Education/ pedagogy, Library & information science

As project management (PM) has become recognized as central to a successful DH project, important resources have been developed to help DH scholars through the entire project lifecycle. However, most of these resources gloss over a tension at the core of implementing PM for DH: that the functionalist models from which PM is drawn is incompatible with the nature of humanistic inquiry. There is an opportunity to adopt a more critical project management approach, and to leverage PM methodologies not only for effectiveness and efficiency, but to establish a project's intellectual, ethical, social and political values.

This workshop explores the notion of critical DH PM by focusing on a foundational best-practice: the project charter. We will introduce the charter as having two-fold importance: 1) it provides the project's organizational infrastructure and 2) it defines and embodies a commitment to a set of values, providing space to negotiate topics like equity and fair labor, accountability, accessibility, sustainability and community practices. Charters are social, setting the tone for collaborative work and establishing the terms on which the project engages the public.

Workshop participants will be provided a template and opportunity to draft and get feedback on their own charters. Additionally, we will discuss and reflect on the charter as a writing and thinking process, engaging with ideas circulating in the DH community such as Ethics of Care/Carework (e.g. Suomela, et al, 2019, and Klein, 2015) and "critical management studies" in the project management community (e.g. Hodgson and Cicmil, 2008).

Creating Digital Collections with Minimal Infrastructure Hands On with CollectionBuilder for Teaching and Exhibits

Olivia Wikle
omwikle@uidaho.edu
University of Idaho, United States of America

Evan Williamson
ewilliamson@uidaho.edu
University of Idaho, United States of America

Devin Becker
dbecker@uidaho.edu
University of Idaho, United States of America

Keywords: static web, digital collections, GitHub Pages, metadata
Topics: Global, English, Contemporary, digital libraries creation, management, and analysis, physical & minimal computing, Education/ pedagogy, Library & information science

Build a digital collection in a day! Using CollectionBuilder-GH, workshop participants will learn fundamental web and DH skills working with CSV data and digital files to generate websites that facilitate browsing and visualizing digital collections in multiple ways. Similar to Omeka, but completely free and highly customizable, CollectionBuilder-GH is an open source project optimized for non-developers and hosted on GitHub Pages. The tool is designed to lower barriers to hosting, development, and access, empowering users to take full ownership over their project while serving it on minimal infrastructure and bandwidth. By the end of this workshop, participants will have gained the knowledge and independence necessary to implement CollectionBuilder in contexts that include creating and disseminating custom research collections or teaching digital exhibits in the classroom.

Culture Analytics Workshop Networks

John Laudun
laudun@louisiana.edu
University of Louisiana, United States of America

Clarisse Clarisse Bardiot
clarisse_bardiot@mac.com
Université Polytechnique Hauts-de-France, France

Mila Oiva
milaoiv@utu.fi
University of Turku, Finland

Keywords: culture analytics, network studies
Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, cultural analytics, data modeling, Humanities computing

Culture Analytics is a collaborative, translational data science that explores culture and cultural interaction as a multi-scale / multi-resolution phenomenon. The macroscopic view, that allows a researcher to move from the microscale of close reading, up through the mesoscales, and on to the macroscale of distant reading, is a hallmark of the field.

Following the 2019 workshop which focused on the use of time series as a way to understand and visualize humanities data, the 2020 workshop is dedicated to how network exploration and analysis can be similarly used to understand and visualize. The questions that focus the workshop are: How can network visualisations provide a distant viewing of data? What are the different steps to build a network? How to read it?

Similar to 2019, the workshop will begin with an introduction to culture analytics and network analysis offered as a framework for the rest of the workshop. Afterwards, there will be two brief presentations of of research examples, one involving a small dataset and another one with big data. This will be followed by a hands-on tutorial with Gephi that will introduce participants to the basics of network exploration and analysis. Gephi is a popular network visualization and analysis tool. It helps scholars investigate empirical phenomena that can be seen as nodes and links, entities and relations. Drawing from graph theory and information design, it helps shifting the focus from the

entities to their relations. It repurposes computational metrics to describe and characterize the topology of empirical networks.

TwinTalks 3 Understanding and Facilitating Collaboration in DH

Steven Krauwer
steven@clarin.eu
CLARIN ERIC

Darja Fišer
darja.fiser@ff.uni-lj.si
University of Ljubljana Jožef Stefan Institute

Keywords: collaboration, cross-disciplinarity, research process, training
Topics: Global, English, 19th Century, 20th Century, Contemporary, project design, organization, management, public humanities collaborations and methods, Humanities computing

Virtual Workshop "TwinTalks 3: Understanding and Facilitating Collaboration in DH"

Workshop organisers:

- Steven Krauwer, CLARIN ERIC / Utrecht University, steven@clarin.eu
- Darja Fišer, CLARIN ERIC / SSHOC / University of Ljubljana, darja.fiser@ff.uni-lj.si

This workshop is a joint initiative of CLARIN ERIC (http://www.clarin.eu) and DARIAH ERIC (http://www.dariah.eu) , and is supported by the SSHOC project (https://sshopencloud.eu/)

Workshop website: https://www.clarin.eu/event/2020/twintalksdh2020

Abstract:

The main objective of the proposed workshop is to get a better understanding of the dynamics on the Digital Humanities work floor when researchers and professionals with different areas of competence work in tandem to solve humanities research questions, and to explore how education and training of both humanities scholars and of technical experts can help making collaboration across disciplines more efficient, effective and (where possible) more creative and innovative. The insights gained should help those involved in the education of humanities scholars with various levels of technical capabilities, as well as technical experts with and/or without the understanding of humanities research to develop better training programmes. To this end we invite (i) submissions by pairs (or teams) consisting of members with different skill sets, as well as (ii) submissions by educators with experience in cross-discipline training. An invited talk at the beginning and a round table discussion at the end will complete the program. Please note that the chosen workshop format with the open call for submissions is in line with common practice in many sub-disciplines of Digital Humanities, and serves to attract contributions from a wide range of disciplines and sub-disciplines of Digital Humanities in a broad sense, including Social Sciences and Cultural Heritage Studies.

Background:

Research in the Humanities in a very broad sense (including Cultural Heritage Studies and Social Sciences) can only truly benefit from new developments in technology if researchers and professionals with different capabilities team up. This situation is very similar to the hard sciences, where research is done in teams working on a specific problem, where everybody brings in his/her specific content and technical expertise and skills. Co-design, co-development and co-creation are the rule rather than the exception, and the proposers want to investigate how this could be facilitated in the disciplines mentioned above. In the rest of this proposal we will use the term *humanities* in the broad sense described above, and *digital humanities* to refer to humanities research using digital data and tools.

The proposed workshop is a follow-up of the first, successful TwinTalks workshop that took place on March 5 2019 at the Nordic DH conference in Copenhagen, and the second TwinTalks workshop to take place at the Nordic DH conference in Riga planned for March 2020, now postponed to October. Whereas in the first workshop the focus was on analysing what happened on the work floor from the perspective of the researchers, and on identifying possible obstacles for collaboration, in the follow-up workshops we have added a new perspective: that of the humanities or digital humanities educator who has experience in bridging the gap between humanities and technical skills. The programme will reflect this.

Objectives:

The main objective of the proposed workshop is to get a better understanding of the dynamics on the Digital Humanities work floor when researchers and professionals with different areas of competence work in tandem to solve humanities research questions, and to explore how education and training of both humanities scholars and technical experts can help making collaboration across disciplines more efficient,

effective and (where possible) more creative and innovative. What does it mean to pose a research question in the humanities that is technically tractable and how does it change our understanding of interpretative work? Conversely, is there room for creativity in the application of technical tools and what are the limitations? How can we ensure that have the same understanding of the research question to be answered and of the technical (im)possibilities?

The insights gained should help those involved in the education of humanities scholars, professionals and technical experts alike to develop better training programmes, tailored towards the needs of a diverse group of potential learners.

Audience:

Researchers, professionals, educators, and Research Infrastructure operators with a special interest in creating the conditions where people with humanities research skills and technical expertise (or both) can fruitfully collaborate in answering humanities research questions.

Originally proposed format (pre-COVID-19):

The programme starts with an invited talk by a prominent speaker, which will set the scene for the rest of the day.

The main component of the workshop programme consists of two types of (submitted) talks:

Twin talks, i.e. talks presented by pairs consisting of someone rooted primarily in humanities research (with a humanities research problem, i.e. not a technical problem or tool) and someone with a more technical background who has contributed technical capabilities to arrive at the answers. Talks consist of three parts, followed by questions from the audience. In the first part the humanities research question is the point of focus, while in the second part the technical aspects of the research are highlighted. In the third part these perspectives come together, as the team describes how the collaboration went, including obstacles that were encountered, and how better training and education could help to make collaboration more efficient and effective.

Teach talks, i.e. talks by people with experience or interesting ideas about how cross-discipline collaboration is or can be addressed in curricula or other training activities.

The final component of the programme, which makes the event a workshop rather than a mini-conference, is a round table discussion between speakers, invited experts and the audience to discuss and analyse the findings of the talks, and to formulate conclusions and recommendations for developing better training programmes.

Revised format:

Due to the coronavirus the DH2020 conference will go virtual, and so will the TwinTalks 3 workshop. This has consequences for the format.

The duration of the virtual workshop will be 2 hours, and it will have an interactive, panel-like format, with short pitches (duration dependent on number of accepted talks). The Twin Talk pitches should briefly describe:

(i) the research problem addressed,

(ii) its solution, including the technical aspects,

(iii) a report on the collaboration experience itself, including obstacles encountered and

(iv) recommendations how better training and education could help to make collaboration more efficient and effective.

The Teach Talks should briefly describe (i) the collaboration settings on which they are based, (ii) the approach adopted, and (iii) recommendations.

After the talks there will be a discussion with all participants to formulate the lessons learned from the presentations, and to identify further steps that could be taken.

Reading material and expected learning outcomes:

The workshop is a follow-up to a successful first Twin Talks workshop at DHN2019 in Copenhagen (https://www.clarin.eu/event/2019/twintalksdhn2019) and a workshop to take place at DNH2020 in Riga (moved to October 2020), where the scene for the discussion of this topic was set.

Reading material:

Participants are requested to read the proceedings of the DHN2019 workshop (http://ceur-ws.org/Vol-2365/) and Mikko Tolonen's invited talk at this workshop "Why humanities research questions should come first? Reflections on different kinds of collaboration in digital history" (see https://docs.google.com/presentation/d/1tXWyJPj47VyUQlHWRrhjQnsXSQuW-bY2QYkvpXYJFUw/edit)

Learning outcomes:

- Recommendations for university and other curricula, and for strategies to disseminate them
- Possible next steps for concrete actions

Revised time schedule:

- Monday, July 6th: Submission deadline for short abstracts
- Friday, July 10th: Notification
- Workshop date and time to be announced later
- Tuesday, September 8th: Submission deadline for extended abstracts

International workshop programme committee (for peer review of submissions and composition of the final programme):

- Bente Maegaard (CLARIN ERIC / University of Copenhagen, Denmark)
- Eleni Gouli (Academy of Athens, Greece)
- Franciska de Jong (CLARIN ERIC / SSHOC / Utrecht University, Netherlands)
- Frank Fischer (DARIAH ERIC / SSHOC / Higher School of Economics, Moscow)
- Frank Uiterwaal (NIOD - KNAW, Netherlands)
- Jennifer Edmond (DARIAH / SSHOC / Trinity College Dublin, Ireland)
- Koenraad De Smedt (University of Bergen, Norway / CLARINO)
- Krister Lindén (University of Helsinki, Finland / FIN-CLARIN)
- Maciej Maryl (Polish Academy of Sciences, Poland)
- Maria Gavrilidou (SSHOC / ILSP - Athena RC, Athens, Greece)
- Radim Hladik (Academy of Sciences, Czech Republic)
- Ulrike Wuttke (RDMO / University of Applied Sciences Potsdam, Germany)

Posters

Effect of Promotion, Rank, and Tenure Guidelines on Open Data Distribution

Talea Anderson
talea.anderson@wsu.edu
Washington State University, United States of America

Keywords: open data
Topics: English, North America, Contemporary, data publishing projects, systems, and methods, data, object, and artefact preservation, Library & information science

Promotion, rank, and tenure (PRT) guidelines have been cited as a key motivation for why and how faculty choose to publish their research. Open-access (OA) advocates have noted in particular that lukewarm or negative portrayals of OA venues in PRT guidelines can result in decreased participation in OA publishing. In response to this concern, some universities have begun to experiment with adding language to PRT guidelines that invites broader participation in publishing venues, including OA publishing. Of interest in this study is the impact of PRT guidelines on the distribution of open data. Specifically, this poster will rely on information gathered from SHARE to consider the prevalence of participation in open data publishing at schools that include more expensive PRT guidelines.

Social media in political engagements in Africa: A study of Twitter use in the intra-party elections of political parties in Ghana

Akwasi Bosompem Boateng
beebeeboateng@gmail.com
North-West University, South Africa

Keywords: Relationship Management, Elections, Political Engagement, Social Media, Twitter
Topics: Africa, English, Contemporary, social media analysis and methods, Communication studies, Political science

The advent of social media is changing the dynamics of political communication, engagements and election campaigns across the world. These platforms extend opportunities for political actors especially political parties, governmenta and citizens to engage directly to build mutually beneficial relationships through engagements ans interactions. Using interviews and content analysis, this article explores "Twittering" in political communication and engagements; how political parties use social media especially during their 2018 intra-party elections in Ghana. It examines how the New Patriotic Party and National Democratic Congress in Ghana use Twitter in intra-party elections. The article gathers data from political parties and their Twitter pages to theorize the political appropriation of social media in Ghana.The findings show that the political parties minimally appropriate Twitter; as they occasionally used it for public information purposes than interactive engagements and mutually beneficial relationships via two-way symmetrical communication. There were insignificant amount of tweet made by the political parties during their intra-party elections in 2018, which indicate that political use of social media in Ghana is still infantile. Therefore, political parties need to improve their use of Twitter and other social media platforms to take maxim advantage of the interactive features and relationship building potentials to garner support from the public and votes in elections. In this regard, political parties require enlightenEd communication strategies and professional political public relations tact for improved use to mature in social media communication.

An online course system easy to make, preserve, and promote critical thinking

Kazushi Ohya
oya-k@tsurumi-u.ac.jp
Tsurumi University, Japan

Keywords: radio lecture system, HTML Imports, Web Components

Topics: Global, English, Contemporary, data publishing projects, systems, and methods, sustainable procedures, systems, and methods, Education/ pedagogy, Media studies

In this poster presentation we will show a new online course consisting of talks and chalks that is easy to make, edit, and preserve, and is a traditional and old-fashioned lecture style, that substantially helps students learn the content spontaneously. The course materials we need to make are only scripts for talks and XML data for slides, and the process we need to undertake separately is just recording the talk. The mechanisms to realize this course are backed with HTML5 and simple codes in JavaScript based on the specification HTML Imports. This lecture system is important not only for an implementation to realize a new teaching/learning channel conveying the knowledge instantly to learners but also to show the evidence for usefullness of HTML Imports as a sign of the existence of users' requirements.

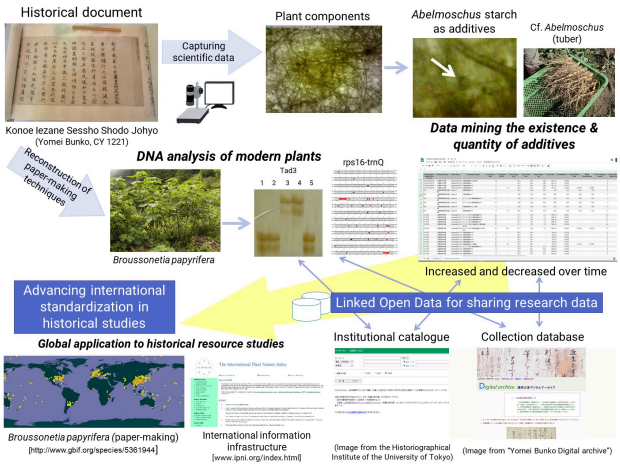
How Do Research Data Develop? International Standardisation of Scientific Data in Historical Studies

Ayako Shibutani
ashibutani@hi.u-tokyo.ac.jp
The University Museum, the University of Tokyo, Japan

Makoto Goto
National Museum of Japanese History, Japan

Keywords: Cultural Heritage Science, Japanese history, linked data, open data, standardization
Topics: Asia, English, 15th-17th Century, 18th Century, 19th Century, digital research infrastructures development and analysis, open access methods, History, History of science

The scientific approach to historical resources creates a synthetic discipline benefitting from open access to data from the humanities and sciences. However, technological challenges exist because of dispersed and heterogeneous resource data. Through data sharing of historical resources, this paper proposes that the establishment of integrated data repositories will capture data provenance and diversity while promoting attribution and acknowledgment of its use. To enhance and accelerate scientific advances in historical studies, we enumerate digital humanities applications to solve the technological and sociological challenges that have limited open access to resource data in the world. We also standardise the scientific methodology of historical materials research using the following approaches: a qualitative analysis focusing on component details to compare our findings with the classifications granted to historical materials in previous studies, and the reconstruction of papermaking methods using DNA analysis.



Improving machine translation literacy to facilitate and enhance scholarly communication

Lynne Bowker
lbowker@uottawa.ca
University of Ottawa, Canada

Keywords: digital literacy; machine translation; machine translation literacy; human-computer-interaction; scholarly communication
Topics: English, North America, Contemporary, artificial intelligence and machine learning, curricular and pedagogical development and analysis, Language acquisition, Translation studies

English is the main language of scholarly communication, but, most researchers are not native English speakers. Contemporary machine translation approaches such as neural machine translation (NMT) are data-driven and use artificial-intelligence-based machine learning techniques; however, such tools rarely produce high quality output of specialized text without human intervention. There is an emerging need for machine translation (MT) literacy among non-Anglpohone students and faculty who must both read and write in English in order to participate fully in the scholarly communication process. We designed and pilot tested a machine translation literacy workshop to help researchers use MT more effectively for scholarly tasks such as: 1) search and discovery of scholarly texts; 2) reading and evaluating scholarly texts; 3) research communication in international teams; and 4) writing for scholarly publishing. Pre- and post-workshop surveys were used to evaluate the success of the workshop and recommend improvements for future iterations.

Faust Times Eighteen: A Network Analysis of Theatre Plays Around the Myth of Faust

Frank Fischer
frank.fischer@dariah.eu
Higher School of Economics, Moscow

Anna Busch
University of Potsdam

Linda-Rabea Heyden
Wikimedia Deutschland e.V. University of Jena

Mark Schwindt
Ruhr University Bochum

Keywords: Faust, drama, plays, literature
Topics: Europe, English, 18th Century, 19th Century, 20th Century, network analysis and graphs theory and application, Literary studies

Alongside the comparative network analysis of larger literary corpora (Algee-Hewitt 2017, Trilcke/ Fischer 2018), there has recently been a trend towards focusing on author-centred subcorpora, such as the oeuvres of Jane Austen (Wade 2017) or Anton Chekhov (Faynberg et al. 2018).Instead of picking out individual authors, we can instead focus on productive literary topoi instead and examine them with the means of network analysis, something that has not been undertaken yet as far as we can see. The abundance of dramas revolving around the Faust myth will serve as an example.The German Drama Corpus (https://dracor.org/ger) currently holds 18 TEI-encoded plays that centre around a Faust character. They range from early plays like Weidmann's "Johann Faust" (1775) to the two parts of Goethe's "Faust" (1808, 1832) and Friedrich Theodor Vischer's "Faust, part III" (1862), but also feature mash-ups like Grabbe's "Don Juan und Faust" (1829) and a version with a female Faust character, Wilhelm Schäfer's "Faustine" (1898).Network analysis enables us to take a comparative macroscopic look at the different structures of these plays. For example, the roles of the devil/sub-devil Mephistopheles and the famulus Wagner can be examined more closely, shedding new light on the structural development of the sujet. It will become clear when and where these characters appear - but also where they are missing. Other types of characters also come to the fore, such as Gretchen or even Faust's parents, who play no role in Goethe's version but who do appear in other plays.Our poster first offers network visualisations that concentrate on the centres, i.e., the constellation directly around the Faust character, especially in the larger plays (Soden's "Doctor Faust" has 62 characters, Julius von Voss's "Faust" has 72 characters, Avenarius' "Faust" has 95 characters, Goethe's first part of "Faust" has 115 and the second part even 189 distinguishable characters/voices). The visualisations are supported by statistical network measures, such as the Betweenness Centrality, Degree and Weighted Degree. In addition to visual evidence, these measures can also address the positioning of the characters surrounding Faust in a new way.The poster will demonstrate how network analytical and quantitative aspects complement existing literary research on the topic (for an overview of the wealth of literary works around Faust cf. Hucke 1992).

Bibliography: Algee-Hewitt, Mark (2017): Distributed Character: Quantitative Models of the English Stage, 1550-1900. In: New Literary History 48(4), pp. 751-782. Johns Hopkins University Press. DOI: https://doi.org/10.1353/nlh.2017.0038

Faynberg, Veronika et al. (2018): Netzwerkanalytischer Blick auf die Dramen Anton Tschechows. In:

DHd2018 Conference Abstracts, University of Cologne, pp. 439. DOI: https://doi.org/10.6084/m9.figshare.6410909

Hucke, Karl-Heinz (1992): Figuren der Unruhe. Faustdichtungen. Tübingen: Niemeyer. Reprint: de Gruyter 2014.

Trilcke, Peer; Fischer, Frank (2018): Literaturwissenschaft als Hackathon. Zur Praxeologie der Digital Literary Studies und ihren epistemischen Dingen. In: Martin Huber, Sybille Krämer (eds.): Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden (= 3rd Special Edition of 'Zeitschrift für digitale Geisteswissenschaften'). DOI: http://dx.doi.org/10.17175/sb003_003

Wade, Karen (2017): Jane Austen's Social Networks. In: The Sea of Books. URL: https://theseaofbooks.com/2017/07/04/jane-austens-social-networks/

Who's Afraid of the Big Bad ? Researching Trends in the Early Gothic Novel using Word Embeddings

Maartje Weenink
maartje.weenink@stu.mmu.ac.uk
Manchester Metropolitan University, United Kingdom

Keywords: word embeddings, gothic, genre formation, sentiment analysis
Topics: Europe, English, 18th Century, 19th Century, database creation, management, and analysis, natural language processing, Linguistics, Literary studies

This poster will present the preliminary results of my PhD project which researches trends in the early Gothic novel using computational methods. A corpus of over 2500 British early Gothic texts has been created and explored using a combination of the analysis of word embeddings and sub-sections of the corpus defined by annotated meta-data. Variations in the embeddings for such sub-corpora demonstrate that various established theories such as the assumed British preoccupation with European national identity in Gothic fiction, the tendency to categorise all Gothic novels as filled with negative sentiment, or the ambiguously defined 'female Gothic', warrant re-evaluation and further exploration assisted by quantitative methods. Notable changes in embeddings for specific datasets such those comprised of texts written at the onset of the Gothic's popularity, or by female authors, are visualised in this poster.

A Flow for Digitizing Japanese Historical Materials and their Long-Term Use

Taizo Yamada
t_yamada@hi.u-tokyo.ac.jp
The University of Tokyo, Japan

Satoshi Inoue
The University of Tokyo, Japan

Keywords: Digitalization, Japanese History, Digital Preservation, OAIS
Topics: Asia, English, 5th-14th Century, 15th-17th Century, data, object, and artefact preservation, digital archiving, Asian studies, History

We formulated a flow and a rule for digitizing historical documents as organizations, not individuals or projects. It is the rule for supplying data, not going to disappear even when projects that create data will end. The flow consists of following processes: the flow has investigating materials and shooting by digital camera, screening and sorting of images, data registration into the database, image registration, security setting. Each process has a responsible person or department in our institution. As a result, images of historical materials of pre-modern Japanese history are acceleratingly concentrated in our storage, and it is growing as "image cloud for Japanese historical material". In order to further develop it, now we have been embedding function of the digital preservation[2] like an Open Archival Information System (OAIS)[3] into our system.

Mind the gap: Filling gaps in cuneiform tablets using Machine Learning Algorithms

Timo Homburg
timo.homburg@gmx.de
Mainz University Of Applied Sciences

Keywords: Cuneiform, Text
Topics: Asia, English, BCE-4th Century, artificial intelligence and machine learning, semantic analysis, Computer science, Linguistics

Introduction

A persisting problem in near eastern studies is the existence of broken cuneiformtablets (listing 1.1). In recent years efforts have been undertaken to 3DScan Maraet al. (2010), to paleographically describe Homburg (2019) and to digitally reconstruct broken fragments Collins et al. (2014, 2017) of cuneiform tablets. However,not always broken fragments can complement each other and often parts of thecuneiform tablet remain destroyed. These fractures or gaps in the cuneiformtablet are not always easy for scholars to fill and take a considerable amount ofinterpretation time on their part. With the emergence of more digitally availablecuneiform text resources, this publication sees an opportunity to investigate ifautocomplete algorithms, based on machine learning and linguistic linked opendata (LLOD) resources Homburg (2017) can be useful in the reconstructionof cuneiform texts. The classification results are to be used to create a epochand language specific recommendation system to fill gaps on cuneiform tablets,therefore assisting cuneiform scholars.

Related Work

Related work has been done in autocompletion systems which face the similarchallenge of anticipating the users input derived from context and other featuresLeung & Zhang (2008), Gikandi (2006), Hyvönen & Mäkelä (2006). Those tech-nologies are heavily relied on in input method engines1 which are powered withdifferent dictionary-based algorithms, but recently Chen et al. (2015), Huanget al. (2018) also with machine learning approaches and neural networks. Inputmethod engines for cuneiform have been developed by Homburg et al. (2015).

Methodology

Following Homburg & Chiarcos (2016) machine learning methods applied areeither based on grammatical rules (POSTagging), dictionary-based methods ex-ploiting (third-party) dictionary resources or statistical approaches using thefollowing types of machine learning features:- Context-dependent features: e.g. for Hidden Markov Model Classifications- Grammatical features derived from POSTaggers- Semantic Features derived from the semantic meaning of surrounding words- Metadata Features e.g. text categorizations- Paleographic Features using PaleoCodage for a subset of manually annotatedtexts Homburg (2019)

Experimental Setup

The effectiveness of the algorithms and features is tested on a corpus of all CDLItexts in ATF which is split in a training and test set. Texts are prepared withrandom gaps for classification and evaluated using the original texts (the goldstandard) on unicode cuneiform and on the respective cuneiform transliterationfor different cuneiform languages (Sumerian, Hittite, Akkadian) and epochs. Theposter features selected peliminary results of the classification and a significanceanalysis of the features for further discussion for improvement. A possible futuregoal could be a shared task to improve classification accuracy similar to thecuneiform language identification challenge Jauhiainen et al. (2019)

Application

Lastly, the poster presents a prototypical application (fig. 1)displaying the results of the machine learning process which is currently in devel-opment. The implementation builds up on the concept of input method enginesHomburg et al. (2015) and will provide a self-learning component.

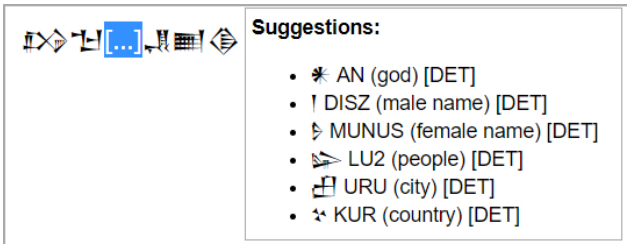


Figure 1: Text Completion Prototype: "If...Enlil". The dictionary knows, that Enlilis a gods name (NE) and is commonly preceded by a determinative character for god??(an), which is suggested in first place to fill the gap. Next likely options are aperson named Enlil (male or female), the people (tribe) of Enlil, or a location

Bibliography

Chen, S., Zhao, H. & Wang, R. (2015), Neural network language model forchinese pinyin input method engine, in 'Proceedings of the 29th Pacific AsiaConference on Language, Information and Computation', pp. 455-461. 2Collins, T., Woolley, S. I., Ch'ng, E., Hernandez-Munoz, L., Gehlken, E., Nash,D., Lewis, A. & Hanes, L. (2017), A virtual 3d cuneiform tablet reconstruc-tion interaction, in 'Proceedings of the 31st British Computer Society HumanComputer Interaction Conference', BCS Learning & Development Ltd., p.

73.1Collins, T., Woolley, S. I., Munoz, L. H., Lewis, A., Ch'ng, E. & Gehlken,E. (2014), Computer-assisted reconstruction of virtual fragmented cuneiformtablets, in '2014 International Conference on Virtual Systems & Multimedia(VSMM)', IEEE, pp. 70-77. 1Gikandi, D. (2006), 'Predictive text computer simplified keyboard with word andphrase auto-completion (plus text-to-speech and a foreign language translationoption)'. US Patent App. 11/308,013. 2Homburg, T. (2017), Postagging and semantic dictionary creation for hittitecuneiform, in 'DH2017'. 1Homburg, T. (2019), Paleo codage - a machine-readable way to describecuneiform characters paleographically, Utrecht, Netherlands.URL: https://dev.clariah.nl/files/dh2019/boa/0259.html 1, 3Homburg, T. & Chiarcos, C. (2016), Word segmentation for akkadian cuneiform,in 'LREC 2016'. 3Homburg, T., Chiarcos, C., Richter, T. & Wicke, D. (2015), 'Learning cuneiformthe modern way'.URL: http://gams.uni-graz.at/o:dhd2015.p.55 2, 4Huang, Y., Li, Z., Zhang, Z. & Zhao, H. (2018), 'Moon ime: neural-based chinesepinyin aided input method with customizable association', Proceedings of ACL2018, System Demonstrations pp. 140-145. 2Hyvönen, E. & Mäkelä, E. (2006), Semantic autocompletion, in 'Asian SemanticWeb Conference', Springer, pp. 739-751. 2Jauhiainen, T., Jauhiainen, H., Alstola, T. & Lindén, K. (2019), 'Language anddialect identification of cuneiform texts', arXiv preprint arXiv:1903.01891 .3.1Leung, B. & Zhang, Q. (2008), 'Providing relevant text auto-completions'. USPatent App. 11/751,121. 2Mara, H., Krömker, S., Jakob, S. & Breuckmann, B. (2010), Gigamesh andgilgamesh:-3d multiscale integral invariant cuneiform character extraction, in'Proceedings of the 11th International conference on Virtual Reality, Archae-ology and Cultural Heritage', Eurographics Association, pp. 131-138. 1

Integrating Intertextual Search into Your Web Application: The Tesserae Intertext Service API

Nozomu Okuda
nozomuok@buffalo.edu
Department of Classics, The University at Buffalo, SUNY

Jeffery Kinnison
jkinniso@nd.edu
Department of Computer Science & Engineering, University of Notre Dame

Neil Coffee
Department of Classics, The University at Buffalo, SUNY

Walter Scheirer
Department of Computer Science & Engineering, University of Notre Dame

Keywords: intertextuality, API, web services, digital classics
Topics: English, North America, BCE-4th Century, 5th-14th Century, Contemporary, software development, systems, analysis and methods, text mining and analysis, Literary studies, Philology

The Tesserae Project presents the TIS API, a REST-based web service that allows partner collections to integrate Tesserae intertext search directly into their web applications.

Visualisation of semantic shifts the case of modal markers

Helena Bermúdez Sabel
helena.bermudezsabel@unil.ch
Université de Lausanne

Francesca Dell'Oro
francesca.delloro@unil.ch
Université de Lausanne

Paola Marongiu
paola.marongiu@unil.ch
Université de Lausanne

Keywords: Latin, Diachronic linguistics, Modality, Semantic maps, Visualization
Topics: Global, Europe, English, BCE-4th Century, Contemporary, data modeling, semantic analysis, Humanities computing, Linguistics

This proposal examines the importance of visual representations to convey semantic shifts while presenting a work in progress concerning the diachronic study of modality in Latin. 1

A visualisation is meant to aid the comprehension of data by taking advantage of our visual perception and its capacity to discern patterns, trends and atypical values (Heer et al. 2010: 59). A visual representation is more accessible, attractive and it can replace complex cognitive calculations. However, selecting the most efficient visualisation can be challenging, especially when conceiving it as a scholarly resource (see Jessop 2008) for the representation of abstract concepts. In our case we would need to:

condensate pages and pages of dictionaries and historical grammars;

add information to previous models to better convey the multidimensionality of modal semantic shifts;

update the traditional visualisation models incorporating motion, color and user interactivity.

The semantic map visualisation method was introduced by Haspelmath (2003)2 to describe and illustrate the multifunctionality patterns of linguistic elements. The semantic map appears as a geometric representation of functions connected together in a semantic space. Semantic maps were employed in various ways, cross-linguistically or on individual languages,3 and synchronically or diachronically.

Van der Auwera and Plungian (1998) apply this resource to visually represent and predict universal patterns of modalisation.4 They build on the single patterns of modalisation for possibility and necessity of the cross-linguistic study by Bybee et al. (1994) (see Fig. 1), complementing them with lexical information from other languages. An overview of the modalisation and grammaticalisation paths is achieved by including pre- and post-modal meanings (Fig. 2). 5

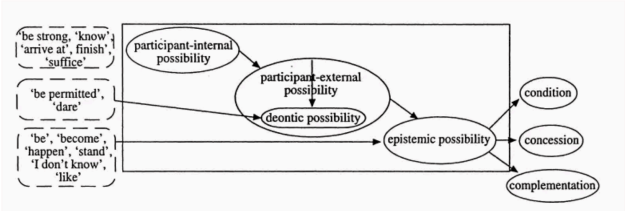


Fig. 1: "To possibility and beyond" (van der Auwera and Plungian 1998: 91).

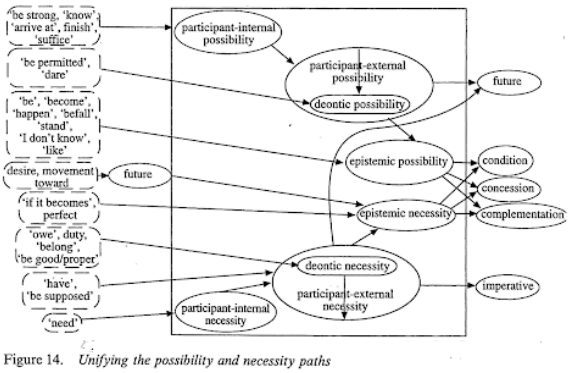


Figure 14. *Unifying the possibility and necessity paths*

Fig. 2: "Unifying the possibility and necessity paths": Example of a semantic map representing the shifts of possibility and necessity (van der Auwera and Plungian 1998: 98).

Our proposal follows this model but our aim is to produce a digital visualisation with these additional features:

Diachrony: Addition of a timeline to visualise when new meanings appear.

Synchrony: Visualisation of coexisting meanings based on position and shape.

Chronology: Addition of the etymological information available for each marker; 6 enrichment of each meaning with its first attestation that is displayed by hovering the mouse over that concept.

Polyfunctionality: Working with empirical data does not imply unambiguous findings, therefore our proposal codifies multiple modal values of the same marker.

Multilingual versions: A map in the source language together with the English translation is available: users can change the language at will.

Legibility: Points 1-5 extend the contents of previous models. To guarantee legibility, certain information is color-coded and other pieces of information will appear by user demand. Colors, for instance, are employed to identify and distinguish pre-, post- and modal meanings. Users can bring out a specific path of modal shift just by clicking on one of the affected senses/steps. A combination of shape and position can render the coexistence of meanings.

We are currently working on the development of the semantic modal maps of some Latin modal markers. These maps were drafted based on the Thesaurus Linguae Latinae (Thesaurusbüro München Internationale Thesaurus-Kommission 1900–), and are susceptible of change against corpus-based evidence. This work in progress is available at our website (http://woposs.unil.ch/semantic-modal-maps.php). Updates refining the visualization and/or adding more semantic maps are frequent. As part

of the future work, we envisage the implementation of additional features: we plan to employ motion to visualise the modal path of a specific marker interacting with its chronological attestations. Also, we are conscious of the challenges that discernity of color entails for people with visual disabilities, thus we want to avoid the use of color as the only visual means of conveying certain information. Therefore, we plan to combine it with other visual cues, like texture. In addition, the size will be customizable and color contrast will be checked. 7

The visualisations currently available were developed using Inkscape 8 to create the basic SVG that will be enhanced by manually including the animation elements with a combination of JavaScript and CSS. The selection of open-source software guarantees an open development. Therefore, not only the results, but all the data and methods will be made publicly available, thus contributing to both the Open Data movement and Public Digital Humanities.

Even if the core of our work concerns Latin modality, the principles and techniques presented can be easily applied to any other language. We claim this model to be versatile: it would be useful to researchers, but also appropriate for vehiculating complex semantic concepts in an educational environment thanks to its readability and immediacy. Besides its possible implementation with languages other than Latin, any type of semantic shift could be visualised following our template. Semantic maps aid in the understanding of meaning so any fields working with natural language, like history or philology, would benefit from our results.

Notes

1This study is part of the SNSF-funded project (SNSF n° 176778) A world of possibilities: Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language (WoPoss). See more information in http://woposs.unil.ch. All the data and code of this project will be made available during the project lifespan (February 2019–January 2023) as open data.

2For earlier conceptualizations of semantic maps see Hjelmslev (1963), Lazard (1981), Anderson (1982).

3For both a single-language approach and a cross-linguistic one see François (2008).

4Semantic maps had been already applied (amongst others) to tense and aspect (Anderson 1982), evidentiality (Anderson 1986), conditionals (Traugott 1985), voice (Croft et al. 1987).

5The model by van der Auwera and Plungian (1998) and Bybee et al. (1994) has been applied by Magni to modal markers in Latin (2005).

6 The etymological information for Latin is taken from the Thesaurus Linguae Latinae (1900–), de Vaan (2011), Ernout-Meillet (1932) and Meiser (1998).

7See https://webaim.org/resources/contrastchecker (consulted on 05/06/2020).

8Available at https://inkscape.org/ (consulted on 05/06/2020).

References

Anderson, Lloyd B. 1982. 'The "Perfect" as a Universal and as a Language-Particular Category'. In Tense-Aspect: Between Semantics & Pragmatics. Amsterdam: Benjamins, 227–64.

--. 1986. 'Evidentials, Paths of Change, and Mental Maps: Typologically Regular Asymmetries'. In Evidentiality: The Linguistic Coding of Epistemology, edited by Wallace L. Chafe and Johanna Nichols, 273–312. Norwood, N.J: Ablex.

Auwera, Johan van der, and Vladimir A. Plungian. 1998. 'Modality's Semantic Map'. Linguistic Typology, 2 (1): 79–124.

Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World. Chicago: University of Chicago Press. https://www.press.uchicago.edu/ucp/books/book/chicago/E/bo3683926.html.

Croft, William, Hava Bat-Zeev Shyldkrot, and Suzanne Kemmer. 1987. 'Diachronic Semantic Processes in the Middle Voice'. In 7th International Conference on Historical Linguistics, 48:179–92.

de Vaan, Michiel. 2011. Etymological Dictionary of Latin. Leiden: Brill. Retrieved Jun 9, 2020, from https://brill.com/view/title/12612

Ernout, Alfred, and Antoine Meillet. 1932. Dictionnaire étymologique de la langue latine: histoire des mots. Paris: Klincksieck.

François, Alexandre. 2008. 'Semantic Maps and the Typology of Colexification: Intertwining Polysemous Networks across Languages'. In From Polysemy to Semantic Change: Towards a Typology of Lexical Semantic Associations, 163–215. Studies in Language Companion 106. https://halshs.archives-ouvertes.fr/halshs-00526845/document.

Haspelmath, Martin. 2003. 'The Geometry of Grammatical Meaning: Semantic Maps and Cross-Linguistic Comparison'. In The New Psychology of Language, 217–48. Psychology press. https://doi.org/10.4324/9781410606921-11.

Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. 2010. 'A Tour through the Visualization Zoo'. Communications of the ACM 53 (6): 59. https://doi.org/10.1145/1743546.1743567.

Hjelmslev, Louis. 1963. Prolegomena to a Theory of Language. University of Wisconsin Press.

Jessop, Martyn. 2008. 'Digital Visualization as a Scholarly Activity'. Literary and Linguistic Computing 23 (3): 281–93. doi:10.1093/llc/fqn016.

Lazard, Gilbert. 1981. 'La Quête Des Universaux Sémantiques En Linguistique'. Actes Sémiotiques (Bulletin) 19: 26–37.

Magni, Elisabetta. 2005. 'Modality's Semantic Maps. An Investigation of Some Latin Modal Forms'. Journal of Latin Linguistics 9 (1): 325–336. https://doi.org/10.1515/joll.2005.9.1.325.

--. 2010. 'Mood and Modality'. In New Perspectives on Historical Latin Syntax. Constituent Syntax: Adverbial Phrases, Adverbs, Mood, Tense, edited by Philip Baldi and Pierluigi Cuzzolin, 2:193–275. Trends in Linguistics. Studies and Monographs 180. Berlin: Mouton de Gruyter.

Meiser, Gerhard. 1998. Historische Laut- und Formenlehre der lateinischen Sprache. Darmstadt: Wissenschaftliche Buchgesellschaft.

Thesaurusbüro München Internationale Thesaurus-Kommission, ed. (1900-). Thesaurus Linguae Latinae. Berlin: De Gruyter.

Traugott, Elizabeth Closs. 1985. 'Conditional Markers'. In Iconicity in Syntax, 289–307.

Analyzing Link Topology to Quantify the Degree of Planned Obsolesce in Online Digital Humanities Projects

Luis Meneses
ldmm@uvic.ca
Electronic Textual Cultures Lab, University of Victoria

Jonathan Martin
King's College London

Richard Furuta
Electronic Textual Cultures Lab, University of Victoria

Ray Siemens
Center for the Study of Digital Libraries, Texas A&M University

Keywords: degradation, abandonment, preservation, online digital humanities projects
Topics: Global, English, Contemporary, data, object, and artefact preservation, digital archiving, Computer science, Humanities computing

Many of the online projects in the digital humanities have an implied planned obsolesce -which means that they will degrade over time once they cease to receive updates in their content and software libraries (Fitzpatrick 2011). We presented papers at Digital Humanities 2017, 2018, and 2019 that explored the abandonment and the average lifespan of online projects in the digital humanities (Meneses and Furuta 2017), contrasted how things have changed over the course of a year (Meneses et al. 2018), and introduced a strategy for preservation by creating standalone software executables (Meneses et al. 2019). However, managing and characterizing the degradation of online digital humanities projects is a complex and pressing problem that demands further analysis.

In this sense "planned obsolescence" is a nuanced designation -as there are many cases of successful projects in digital humanities that are shifting their focus from active development to data management (for example: http://cervantes.dh.tamu.edu). These are cases where a project's online presence has not received updates for some time but its online tools are stable and continue to be accessed by its users. However, if updates are not applied to the infrastructure or content of a project over time web requests will eventually start generating errors on the server or the client -affecting the overall user experience (Nowviskie and Porter 2010). These are examples of why the rules for traditional resources do not fully apply and new metrics are needed to identify issues concerning online projects in the digital humanities.

In this study we dive deeper into exploring the distinctive signs of abandonment to quantify the planned obsolesce of online digital humanities projects. In our workflow, we use each project included in the Book of Abstracts that is published after each Digital Humanities conference from 2006 to 2019. We then proceed to periodically create a set of WARC files for each project, which are processed using Python (van Rossum 1995) and Apache Spark (Apache Software Foundation 2017) to statistically analyze the retrieved HTTP response codes, number of redirects, DNS metadata and detailed examination of the contents and links returned by traversing the base node. This combination of metrics and techniques has allowed us to assess the degree of change of a project over time. As one of the results from our 2019 presentation,

we claimed that the most important signature for degradation comes from the assessing the validity and overall health of the topology of links in a project. Thus, the focus of our study is analyzing this key signature.

We acknowledge that research on the preservation of projects in the digital humanities is also carried out by other groups (Larrousse and Marchand 2019) (Arneil, Holmes, and Newton 2019). However, our study is different as it focuses on two points: first, identifying the signals of abandoned projects using computational methods; and second, quantifying their degree of abandonment. In the end, we intend this study to be a step forward towards better preservation strategies for the planned obsolesce of online digital humanities projects.

References

Apache Software Foundation. 2017. "Apache Spark: Lightning-Fast Cluster Computing." April 11. http://spark.apache.org.

Arneil, Stewart, Martin Holmes, and Greg Newton. 2019. "Project Endings: Early Impressions From Our Recent Survey On Project Longevity In DH." presented at the Digital Humanities 2019, Utrecht, The Netherlands, July 9.

Fitzpatrick, Kathleen. 2011. Planned Obsolescence: Publishing, Technology, and the Future of the Academy. New York: NYU Press.

Larrousse, Nicolas, and Joel Marchand. 2019. "A Techno-Human Mesh for Humanities in France: Dealing with Preservation Complexity." presented at the Digital Humanities 2019, Utrecht, The Netherlands, July 9.

Meneses, Luis, and Richard Furuta. 2017. "Shelf Life: Identifying the Abandonment of Online Digital Humanities Projects." presented at the Digital Humanities 2017, Montreal, Canada, August 8.

Meneses, Luis, Jonathan Martin, Richard Furuta, and Ray Siemens. 2018. "Part Deux: Exploring the Signs of Abandonment of Online Digital Humanities Projects." presented at the Digital Humanities 2018, Mexico City, June 26.

---. 2019. "A Framework to Quantify the Signs of Abandonment in Online Digital Humanities Projects." presented at the Digital Humanities 2019, Utrecht, The Netherlands, July 9.

Nowviskie, Bethany, and Dot Porter. 2010. "The Graceful Degradation Survey: Managing Digital Humanities Projects Through Times of Transition and Decline." presented at the 2010 Conference of the Alliance of Digital Humanities Organizations. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-722.html.

van Rossum, Guido. 1995. "Python Tutorial, Technical Report CS-R9526." Amsterdam: Centrum voor Wikunde en Informatica (CWI). https://ir.cwi.nl/pub/5007/05007D.pdf.

From Metadata to linked Open Data and Wikidata: Yemenite Hebrew Manuscripts and Wikidata

Yitzchak Miller
Isaac.Miller@biu.ac.il
Bar Ilan University, Israel

Gila Prebor
gila.prebor@biu.ac.il
Bar Ilan University, Israel

Keywords: linked Open Data, Wikidata, Hebrew Manuscripts, Metadata
Topics: Asia, English, Contemporary, linked (open) data, manuscripts description, representation, and analysis, Humanities computing, Library & information science

The Hebrew Manuscripts catalogue in the National Library of Israel provides metadata for most Hebrew manuscripts in the world. Traditionally, library catalogues have served as a tool to manage library collections causing library catalogues to be data silos. In order to break down these metadata silos, information must be accessible and free. The semantic web, and in particular, linked open data, are initiatives that can turn library catalogues into a real part of the Internet. In order to explore the potential of linked data to this type of data we plan to convert a small part of the catalogue's metadata of Hebrew Manuscripts in the NLI to Wikidata and create Wikidata items for each of the chosen manuscripts. We hope this will lead to an enrichment of the data and easy access to tools for querying and visualizing the collection. As a case study we have chosen Yemenite script manuscripts.

Enhanced Stand-off TEI Annotation with StoReCo A generic approach with the use of RDF

Richard Hörmann
richard.hoermann@sbg.ac.at
University of Salzburg, Austria

Daniel Schlager
daniel.schlager@sbg.ac.at
University of Salzburg, Austria

Keywords: Stand off markup RDF Annotation
Topics: Europe, English, Contemporary, data modeling, text encoding and markup language creation, deployment, and analysis, Humanities computing, Informatics

TEI-XML is the standard in digital text editing. The classic inline annotation with TEI quickly reaches its limits when it comes to complex annotations. In order to realize DH projects with such requirements on the dhPLUS platform currently developed at the University of Salzburg, a special form of stand-off markup is presented: It saves the original TEI file, converts the markup into RDF instructions, keeps the text and links the triples to an unlimited number of annotation levels. A unique identifier connects the different stages and facilitates presentation and search processes.

Colorization of Illustrations in Charles Dickens' Novels Using Deep Learning

Hoyeol Kim
elibooklover@gmail.com
Texas A&M University, United States of America

Maura Ives
m-ives@tamu.edu
Texas A&M University, United States of America

Keywords: Colorization, Deep Learning, Illustration, Victorian Literature, Charles Dickens
Topics: Global, Europe, English, North America, 19th Century, artificial intelligence and machine learning, digital art production and analysis, Education/ pedagogy, Literary studies

Charles Dickens dedicated himself to the development of Victorian visual culture by actively employing illustrations in his fiction. For Dickens, illustrations were a way to attract readers and boost sales, as demonstrated by the commercial success of The Pickwick Papers (1837), as well as a crucial element of artistic expression. All but two of Dickens's novels were illustrated, and his involvement in every stage of the illustration process was well documented. Dickens collaborated closely with his illustrators by providing detailed instructions: he provided specific colors for his illustrations, although the illustrations would be printed in black and white, and he intentionally positioned each illustration in a specific location in his serials in order to communicate details and emotions effectively with his readers. Scholars of Dickens's works generally understand the illustrations to be integral to the text.

Given the expense of printing illustrations in color, most of the illustrations in Dickens's work were printed in black and white. However, the hand-colored illustrations in A Christmas Carol (1843) demonstrate both Dickens's interest in providing color when he could, and the significance of color in interpreting design elements as well as shaping interpretation of Dickens's text. Because colorizing Dickens's illustrations has the potential for enhancing the reader's understanding of the text, and for opening up new interpretive possibilities, it has pedagogical implications that we wish to explore.

Our project is the first deep learning colorization venture in Victorian era media. We will colorize all of the illustrations in several of Dickens's major novels, using the pix2pix model based on cGANs with Kim's Victorian400 dataset as a research method. The Victorian400 dataset, published as an open data source, is a collection of colorful illustrations painted with nineteenth-century palettes. By using Victorian paintings and hand colored illustrations as a training set, our research methods make it possible for Dickens's illustrations to be viewed with a Victorian color palette, approximating the color choices that a contemporary audience would have expected. We will present two groups of students with both the colorized and black and white illustrations and ask them to evaluate each in terms of design and in terms of their effect on their engagement with the text. Colorization with deep

learning carries possibilities of misrepresentation or misinterpretation, which can be used to clarify students' preconceptions about the nature or role of illustration and to spur creative responses to the text; it can also provide interest, anticipation, and imagination for readers. As a pedagogical tool, we hope that colorization using deep learning will improve students' ability to think critically about the intersection of text and image while promoting deeper understanding and interest in Dickens's works, and lay the groundwork for experimentation with deep learning colorization of other illustrated works by Victorian authors.

Analysis and Visualisation of Complex Familial Relationships in Greek Mythology

Yaya Chenyue Lu
yaya.lu@anu.edu.au
Research School of Computer Science,
The Australian National University

Ben Swift
Research School of Computer Science,
The Australian National University

Greta Hawes
Research School of Humanities & the Arts,
The Australian National University

Keywords: Greek mythology, genealogy, graphic design, relationship analysis, graphing algorithms
Topics: Global, Europe, English, BCE-4th Century, Contemporary, digital biography, personography, and prosopography, Interface design, development, and analysis, Humanities computing, Literary studies

Greek Mythology contains numerous idiosyncratic familial relationships that pose unusual challenges for systematic analysis and visualisation. This project examines these idiosyncracies by using a subset of the Greek Mythology data collected according to the MANTO ontology developed by Dr Greta Hawes. This includes the design and development of a user-friendly web-based tool and graph layout that visualises and explores these complex relationships in the context of the digital humanities. In the process, it hopes to generate new avenues of research for data visualisation of complex linked open data structures, and raise societal awareness for unconventional "family" and "relationship" types.

From Millions of Words to a Single Phrase: Examination of the TXM Software in Hebrew

Maxim Lengo
maximlengo@gmail.com
Bar-Ilan University, Israel

Keywords: Textometrical Research, TXM software, Online Social Networks
Topics: Global, Europe, English, North America, Contemporary, database creation, management, and analysis, social media analysis and methods, Communication studies, Linguistics

Following the technological development during the second half of the 20th Century, linguistic researchers have begun for the first time to analyze corpora of big data. One research methodology, which was developed for lexicometry and text statistical analysis, is Textometry. I.e., the attempt to combine various statistical analysis techniques, such as factorial correspondence analysis (Benzécri, 1977) and hierarchical ascendant classification (Ward Jr, 1963), with full-text search techniques such as kwic concordances (Luhn, 1960), in order to trace the precise original editorial context of any textual event participating to the analysis.

The TXM software (Heiden, 2010), which was developed in France as a modular platform of a new generation of textometrical research and which I adapted to Hebrew, gives the ability to analyze a large corpus of texts as in my research, by using tools and methods based on linguistics and discourse analysis, that is, decomposing the text into factors and elements, carrying out statistical analysis, identifying the hidden social patterns, and then restoring the corpus to its original mode.

As a Ph.D. student in the discipline of Social Sciences, my research focuses on the image repair theory (Benoit, 2015) as an ensemble of strategies such

as evading responsibility and reducing offensiveness, used by individuals, organizations and groups in order to repair their image during times of crisis. It examines the ways in which rhetorical measures are used in online verbal exchanges among users of the online social networks who attempt to repair their personal image. To achieve my goal, I use the TXM software to analyze a corpus of more than eight million words in 365 Facebook posts, which were published by the Israeli Prime Minister Benjamin Netanyahu during his current affairs, and more than 285,000 comments, made by the users. Netanyahu's Affairs are four police investigations in which he is involved as a suspect or has given a testimony.

During the poster session, I will present a review of some tools I used with the TXM software during the digitized analysis process in my Ph.D. research and the way they allowed me to recognize the following key phrase used by Netanyahu: "They have the media, we have you". Among these tools are Progression, which is the frequency of occurrence of one term throughout the text corpus; Co-occurrence, which is the frequency of occurrence of two terms in a text corpus alongside each other in a certain order; and Specificity, which is the score a term is given based on its occurrence in the corresponding part of the corpus relative to the one in the entire corpus, and indicates whether it is overused, underused or useless.

Bibliography

Benoit, W. L. (2015). Accounts, excuses, and apologies: Image repair theory and research. Albany, New York: State University of New York Press.

Benzécri, J. P. (1977). Sur l'analyse des tableaux binaires associés à une correspondance multiple. Cahiers de l'Analyse des Données, 2(1), 55-71.

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In R. Otoguro, Ishikawa, K., Umemoto, H., Yoshimoto, K., Harada, Y. (Eds.), 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24 (Pp. 389-398). Waseda University, Sendai, Japan: Institute for Digital Enhancement of Cognitive Development.

Luhn, H. P. (1960). Key word?in?context index for technical literature (kwic index). American Documentation, 11(4), 288-295.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), 236-244.

Xtralingua: An open-source tool for extracting quantitative text profiles

Panagiotis Papantonakis
Hellenic OCR Team

Fotios Fitsilis
Hellenic Parliament

Sotiris Leventis
Hypernetica

George Mikros
gmikros@gmail.com
Hamad Bin Khalifa University, Qatar

Keywords: quantitative text analysis, quantitative text profiles, Hellenic OCR Team, Google Summer of Code, Open Science
Topics: Global, English, Contemporary, natural language processing, text mining and analysis, Humanities computing, Linguistics

The aim of this poster is to present a novel tool for extracting quantitative text profiles from corpora using a friendly Graphical User Interface. Xtralingua is a software that incorporates over 60 specialized quantitative text analysis measurements including text readability and lexical diversity indices as well as specialized measurements in the text inspired by theoretical work done in the area of Quantitative Linguistics. The tool is open-source and can be further enriched with custom quantitative text indices that the users can add using a scripting language. Xtralingua offers researchers with no specialized technical skills the ability to quickly extract rich quantitative text profiles for further processing. Moreover, it is easy to operate and can support both research and teaching needs in a variety of DH topics.

Les Misérables & the CITE Architecture: A Publication and Toolkit

Christopher William Blackwell
cwblackwell@gmail.com
Furman University, United States of America

William Blackwell
wblackwell98@gmail.com
Furman University, United States of America

Max Norman
max.norman@furman.edu
Furman University, United States of America

Keywords: Hugo, Citation, N-Gram, Scala, Paris
Topics: Europe, English, French, 19th Century, scholarly editing and editions development, analysis, and methods, text mining and analysis, Humanities computing, Literary studies

Our poster will present and link to a publication and toolkit for working with Victor Hugo's Les Misérables, in French and English, using the CITE Architecture. The published data will include CTS-Compliant texts in French and English, and programmatically derived versions of those texts: TEI-XML, HTML, stop-words removed (useful for Topic Modelling), lemmatized (stemmed) editions, vocabulary lists, contextualized concordance, and a web-based translation-alignment tool.

The deliverable is not only a very rich deluxe, bilingual edition of the novel, but the documented scripts used to take a CITE/CTS text and transform it for different presentations and analyses.

"To be or not to be central" - On the Stability of Network Centrality Measures in Shakespeare's "Hamlet"

Anastasia Klimashevskaia
klimashevskaya.anastasia@gmail.com
Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria

Bernhard C. Geiger
Know-Center GmbH, Graz, Austria

Martin Hagmüller
Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria

Denis Helic
Institute of Interactive Systems and Data Science, Graz University of Technology, Graz

Frank Fischer
Higher School of Economics, Moscow

Keywords: literary network analysis
Topics: Europe, English, 15th-17th Century, Contemporary, artificial intelligence and machine learning, network analysis and graphs theory and application, Informatics, Literary studies

Introduction

Centrality measures derived from character networks can be used to detect the main characters in a play. For example, previous research has shown that characters with high network centrality typically perform the majority of speech acts and appear in most of the scenes (Fischer, Trilcke, Kittel, Milling, & Skorinkin, 2018). However, one can extract character networks from plays in various ways: Close reading may omit minor characters like attendants or servants, e.g., (Moretti, 2011), while distant reading (e.g., parsing an XML file) may include aggregate characters like "All", "Both Lords", or similar. Furthermore, the networks may display either implicit or explicit connections, depending on whether we connect characters because they appear in the same scene or because they are directly addressing each other, respectively. Thus, as adding more characters or connections to the network affects centralities and other network measures, the interpretation of both qualitative and quantitative aspects of characternetworks depends on the extraction method. In this work we are concerned with the specific question whether details of the textual source and the extraction method, such as adding minor or aggregate characters, make the main characters less "central". A negative answer to this question would provide us with a further evidence for the validity of automated literary network analysis.

Approach

We analyse six versions of the character network of Shakespeare's "Hamlet". All networks were extracted via close or distant reading from different XML or text sources (see Figure 1) and analysed with NetworkX (Python). For each network, we compute four different centrality measures (closeness, betweenness, degree, and eigenvector centrality). Subsequently, for each centrality measure, we rank the 26 characters common in all networks and compare character ranks in different networks by computing their Spearman rank correlation.

Dataset	Implicit/Explicit Close/Distant	# of characters	# of Edges	Avg. DEGREE	Diam.	Avg. Path Length
Dracor (Fischer, Trilcke, & Orekhov, Drama Corpora Project, 2019)	IMPLICIT/DISTANT	38	220	11.57	3	1.74
Ezlinavis (Milling & Fischer, 2019)	IMPLICIT/DISTANT	34	192	11.29	3	1.70
Grandjean (Grandjean, 2015)	IMPLICIT/DISTANT	37	263	14.22	3	1.63
Haworth (Haworth, 2019)	IMPLICIT/DISTANT	35	217	12.4	3	1.67
Own XML Parser	EXPLICIT/DISTANT	35	74	4.23	4	2.3
Moretti (Moretti, 2011)	EXPLICIT/CLOSE	28	54	3.86	4	2.29

Figure 1. Basic statistics of the character networks

Observations, Conclusion, and Outlook

The networks including implicit connections are denser than those with only explicit connections (cf. Figure 1). This yields different centrality ranks including the most important characters (cf. Figure 2). For example, Horatio has many more implicit connections and connections to minor characters, which makes him the character with the highest degree centrality in the Haworth network. In the Moretti network, which contains only explicit connections, Hamlet has the highest degree centrality.

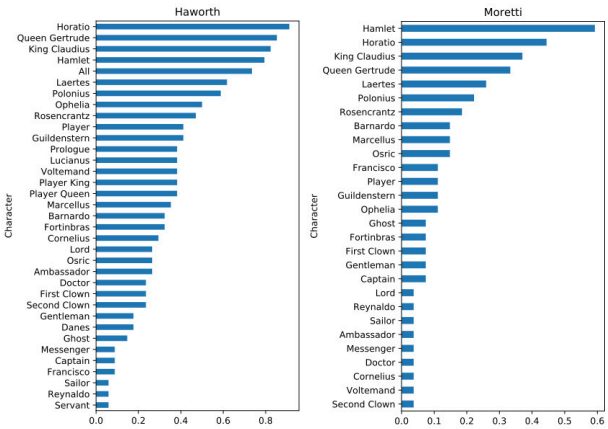


Figure 2. Degree centralities for the Haworth (implicit connections, distant reading) and Moretti (explicit connections, close reading) networks

Despite such individual differences, the groups of main characters derived from different networks exhibit relatively stable rankings, cf. (Fischer, Trilcke, Kittel, Milling, & Skorinkin, 2018). In contrast, rankings for minor characters tend to differ significantly (see Figure 3). Therefore, for detecting the group of main characters, the details of the network extraction method do not have a significant effect, at least in the datasets we consider. In future work we aim to validate the generality of this claim by considering larger corpora of dramatic plays.

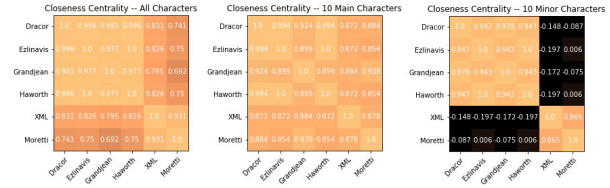


Figure 3. Heatmaps depicting the rank correlation between closeness centralities derived from different networks for all, the 10 most important, and the 10 least important characters

Finally, we outline some further observations about different centrality measures: In our datasets, degree centrality is the most robust, exhibiting high rank correlation for all considered sets of characters. In contrast, eigenvector centrality has the widest range of rank correlations suggesting its high sensitivity with respect to the network structure.

Acknowledgements

The authors gratefully acknowledge permissions to use material from Martin Grandjean and Roger Haworth. The work was funded by the HRSM project "KONDE - Kompetenznetzwerk Digitale Edition". The work of Bernhard C. Geiger was partially funded by the Austrian Academy of Sciences within the go!digital Next Generation project "DiSpecs" (GDNG_2018-046_DiSpecs). The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under die auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Digital and Economic Affairs, and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

Fischer, F., Trilcke, P., & Orekhov, B. (2019). Drama Corpora Project. Retrieved from github.com/dracor-org/shakedracorFischer, F., Trilcke, P., Kittel, C., Milling, C., & Skorinkin, D. (2018). To Catch a Protagonist: Quantitative Dominance Relations in German Language Drama (1730-1930). Digital Humanities 2018 Puentes-Bridges. Book

of Abstracts, 193-201.Grandjean, M. (2015). Network visualization: mapping Shakespeare's tragedies. Retrieved from martingrandjean.ch/network-visualization-shakespeare/Haworth, R. W. (2019). Who appears where matrices and reader's versions for Shakespeare plays. Retrieved from http://rhaworth.net/shakes/shaxref.php?csv=on&targ=on&dnl=on&p=menu#dummyMilling, C., & Fischer, F. (2019). Easy Linavis. Retrieved from https://ezlinavis.dracor.org/Moretti, F. (2011, March). Network Theory, Plot Analysis. New Left Review(68), pp. 80-102.

Representing Genres with distinctive Words

Fotis Jannidis
Würzburg University

Leonard Konle
leonard.konle@uni-wuerzburg.de
Würzburg University

Peter Leinen
German National Library

Keywords: Genre, Summarization, Modeling, Machine Learning, Dime Novel
Topics: Europe, English, 20th Century, Contemporary, data modeling, text mining and analysis, Computer science, Literary studies

We use Burrows Zeta, Semantic Zeta, LDA, Recursive Feature Elimination and Attention Networks to generate lists of distinctive Words for literary genres in pulp fiction and evaluate their quality with a reader survey and document clustering.

The DARIAH ELDAH consent form wizard

Vanessa Hannesschläger
vanessa.hannesschlaeger@oeaw.ac.at
Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Austria

Koraljka Kuzman Šlogar
koraljka@ief.hr
Institute of Ethnology and Folklore Research, Zagreb, Croatia

Walter Scholger
walter.scholger@uni-graz.at
Centre for Information Modelling, University of Graz, Austria

Keywords: Consent Forms, Legal Consent, Legal Issues, Data Privacy
Topics: Global, Europe, English, Contemporary, copyright, licensing, and permissions standards, systems, and processes, digital access, privacy, and ethics analysis, Humanities computing, Law and legal studies

In this poster, we would like to introduce the "consent form wizard" developed by the DARIAH-EU working group "Ethics and Legality in Digital Arts and Humanities" (ELDAH). This tool is currently under development and will be finalized and available in time for the DH2020 conference. It will enable digital scholars as well as the wider research infrastructure community to quickly and easily obtain a standardized consent form that is conforming to the obligations and regulations of the European Union General Data Protection regulation (GDPR) and therefore legally valid in all of the European Union.

Depending on the context for which consent is obtained (e.g. use of images of people at academic events, processing of information shared in surveys, collection / processing of personal data), the wizard users will receive a consent form tailored to their specific needs after answering a series of questions. The different scenarios are the result of discussions with participants of the DH2019 in Utrecht and the DARIAH Annual Meeting 2019 in Warsaw and hence driven by the needs of the research community. Additional meetings with stakeholders from cultural heritage organisations (e.g. ICARUS) and the DARIAH-EU community will ensure a critical evaluation as well as a pragmatic, user-friendly implementation of the tool.

It will be inspired by the code of the CLARIN LINDAT Public License Selector (cf. Kamocki et al. 2016) which offers the same service for license selection that we would like to develop for obtaining a consent form. Software developers, legal experts and (digital) humanities researchers will cooperate in the development of this tool, which - while legally addressing the strict data protection regulations of the

European Union - will also be of use for the international research community outside Europe, since the consent forms will provide a best-practice template for ethical research conduct when processing personal data, and hence address the increasingly prominent topic of ethical research practice and scientific behaviour, especially in a largely digital, Internet-based research context (cf. Markham & Buchanan 2012, McKee & Porter 2009).

The presentation of the poster at the DH2020 conference will allow us to present the wizard to colleagues outside our European network, discuss their needs and experiences with the protection of personal data and hence, if appropriate, adapt the wizard to reflect a global rather than European common perspective on research ethics and the processing of personal data.

References

Ethics and Legality in Digital Arts and Humanities (DARIAH ELDAH). Blog. https://eldah.hypotheses.org/

International Centre for Archival Research (ICARUS). https://icar-us.eu/

Kamocki, Pawe?, Pavel Stranák, and Michal Sedlák (2016). "The Public License Selector: Making Open Licensing Easier." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, et al., 2533-2538. Paris: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/pdf/880_Paper.pdf

Markham, A., Buchanan, E. (2012): Ethical Decision-Making and Internet Research Recommendations from the AoIR Ethics Working Committee. http://aoir.org/reports/ethics2.pdf

McKee, H. A., Porter, J. E. (2009). The Ethics of Internet Research: A Rhetorical, Case-based Process.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection regulation GDPR). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016R0679

DH Teaching Activities as a Resource for Research Accessing the DH Course Registry Data via an API

Tanja Wissik
tanja.wissik@oeaw.ac.at
Austrian Academy of Sciences, Austria

Hendrik Schmeer
Austrian Academy of Sciences, Austria CLARIN ERIC

Frank Fischer
DARIAH EU

Franciska de Jong
CLARIN ERIC

Walter Scholger
University of Graz, Austria

Leon Wessels
CLARIN ERIC

Keywords: API, DH teaching, data access, data publishing
Topics: Global, Europe, English, Contemporary, curricular and pedagogical development and analysis, data publishing projects, systems, and methods, Education/ pedagogy, Humanities computing

The presentation of the DH Course Registry and, more specifically, the possible use of the data through its API at DH2020 is envisioned to attract an increase of attention and participation by colleagues from beyond the European DH research and teaching community and constitutes an effort towards a global communication on DH pedagogy.

Mythologies: A Sonification of Concept Formation in Relation to the Idea of a University

David Michael Berry
d.m.berry@sussex.ac.uk
University of Sussex, United Kingdom

Barnaby Thorn
barnabythorn@gmail.com
University of Sussex, United Kingdom

Keywords: Sonification, Sound, University, History of the University, Critical Digital Humanities
Topics: Global, Europe, English, North America, 19th Century, 20th Century, Contemporary, cultural analytics, music and sound digitization, encoding, and analysis, Cultural studies, Media studies

Mythologies is a sonic composition drawing from work researched through the project "Reassembling the University" in 2017-18 and the underlying "Idea of a University" from approximately 1840 to the present day. This looked at how digital transformations were remaking what we think of as knowledge formed in and through the university over time. Part of the outcomes of that project was funding for the ØxØ project as a sonic installation. This included looking at how concepts, organisation and computation are formed into a new constellation of power and knowledge. This work examines how these systems are legitimated and the orders of justification around them, together with the potential of exemplar concepts such as "utility" for providing immanent critique and the space for practices of critical reason. In this research we have started to describe this new form of university as a data-intensive university, one that is increasingly challenged by a data-centric economy and polity. Associational modes of thinking are particularly encouraged through concepts of streaming, surfing, or algorithmic reverie. The possibility of thought is thereby prescribed by a net of software that surrounds and envelopes the mind, creating a new topology of attention. In this research the university, in its work in the creation of critical reason is argued to be a key, if not the key, institution to contest these developments. We explore that these shifting social, political and economic forces that call for a new distinctive role for the university through a method of sonification that works through concepts immanent to the university represented through sound. But the research also raises new questions for the university: How do the academic disciplines respond, what are the implications for their teaching and research? What are the implications for the wider university? What will be the pattern or shape of this new university? It has been argued that the chapel and the library are no longer the centre of a university, so where is the centre of the university in a data-intensive age? This poster documents the processes and conceptual work undertaken in this research and the eventual final audio and visual materials that have been produced. Most notably the poster displays how sonification can highlight and explore through sonic means the conceptual changes in research data and thereby contribute to public digital humanities through sonic forms of communication. In addition to the poster and material copies, which have been published as tape cassette as well as digital versions, audio versions of the entire piece will be shown with the poster display available through headphones and embedded within the poster format using an experimental "sound button" technology. The sonic work is 33 minutes (made up of five distinct but intertextual compositions: 1.Edge 05:33, 2. Core [Overture] 05:12, 3. Cloud 05:17, 4. Mythologies 04:13, 5. Core [Original] 11:03) and was produced and released on 4th October 2019.

For more information visit: http://www.self-titledmag.com/oxo-feature-mythologies-ep/

INTEGRATING AN ANNOTATION TOOL INTO A VALIDATION PROCESS THE CASE OF READ-IT.

Francois Vignale
francois.vignale@univ-lemans.fr
Le Mans Université, France

Alex Hebing
Utrecht Universiteit, Netherlands

Julian Gonggrijp
Utrecht Universiteit, Netherlands

Quentin Morcrette
Le Mans Université, France

Keywords: annotation, interface, UX design
Topics: Global, Europe, English, 20th Century, Contemporary, annotation structures, systems, and methods, user experience design and analysis, Book and print history, Cultural studies

This poster aims to present a central element of the strategy for validating the concepts and data model of the READ-IT project (https://readit-project.eu), whose purpose is to explore reading experiences scattered in digital sources to answer such questions as "why" and "how" people read. But it still needed validation through an iterative process. In this regard, an specific annotation tool was developed. By annotating very specific characteristics of the act of reading, our tool has a triple function and takes place into a wider framework: 1) to validate a model and the ontology derived from it; 2) to produce vocabulary for specific thesauri; 3) to help answer the fundamental HSS research question of why and how people read through time and space. This tool, whose design is based on UX design approach, allows full integration of any ontology, as long as the ontology is RDF compliant.

Beside and Beyond: Visualising the Paradata and Metadata of Digitised Historical Newspapers with SKOS and LOD

Emily Bell
e.bell@lboro.ac.uk
Loughborough University, United Kingdom

M. H. Beals
Loughborough University, United Kingdom

Keywords: newspapers, metadata, linked open data, SKOS
Topics: Global, English, 19th Century, 20th Century, Contemporary, digitization (2D & 3D), metadata standards, systems, and methods, History, Library & information science

Over the past thirty years, national libraries, universities and commercial publishers around the world have preserved and made available hundreds of millions of pages of historical newspapers through mass digitisation, currently releasing over one million new pages per month. These have become vital resources not only for humanities researchers but also for journalists, politicians, schools, and the general public.[1] However, research conducted by the multinational Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914 project has shown that the collections created by digitisation programmes have not always accurately represented the cultural histories of these publications and their producers, especially the relationships with peoples and institutions in other regions;[2] the very creation of national collections obscures the reality that global news exchanges were central to the nineteenth-century press.[3] The role and repercussions of the individual is equally seen in the digitisation process. On the one hand, individuals from a variety of backgrounds, and with a diverse range of remits, create internal policies and processes that shape how newspapers are stored digitally. One the other, individual end users, with different interests and competencies, must work across diverse collections in order to build as full a picture as possible of global information networks, leading to undocumented irregularities and inconsistencies in their research samples.

This poster will explore our efforts to bridge the interoperability gap between those creating the authoritative, standardised metadata for these collections and the end users attempting to create historical and other narratives through the use of these materials. It has two objectives:
To clarify the decision-making processes behind the structure and creation of digital archives, specifically selection and metadata; To offer historically informed guidelines and models for archives and end-users, to enhance the quality of engagement and development going forward.

Our work aims to enable researchers from a variety of disciplinary backgrounds to break through the barriers between siloed collections as well as provide historically-informed principles for archivists and digitisers to consider when implementing their metadata standards.[4] The poster will present visualisations of existing metadata standards for digitised newspapers mapped to researcher-focused ontologies, based on supplementary information and context (the paratext)[5] and developed by the

Oceanic Exchanges project using the Simple Knowledge Organization System (SKOS). It will demonstrate how the structures of digital and physical newspaper archives can be explored and manipulated to better inform humanities research questions using semantic web technologies.[6] These visualisations will link to a further resource developed by the Oceanic Exchanges project: The Atlas of Digitised Newspapers, which provides a comparative history of key digitised newspaper collections, including their selection processes, digital infrastructures (including Xpaths and technical definitions) and licensing requirements. These elements-the ontology visualisations and the Atlas -will provide a thoroughly annotated controlled vocabulary designed to be used across disciplines, inside and outside the academy, and provide the basis for further digital humanities research using digitised historical collections.[7]

[1] Paul Gooding, Historic Newspapers in the Digital Age: "Search All About It!" (Basingstoke: Palgrave, 2017): 172; James Mussell, The Nineteenth-Century Press in the Digital Age (Basingstoke: Palgrave, 2012): 1.

[2] In conjunction with partners at institutions including Northeastern University, Universität Stuttgart, Universidad Nacional Autónoma de México, Universiteit Utrecht, Turun Yliopisto and University College London; see https://oceanicexchanges.org/team/.

[3] M. H. Beals, "Transnational Exchanges" in David Finkelstein, ed., The Edinburgh History of the British and Irish Press, vol. 2 (Edinburgh: Edinburgh University Press, 2020): 240-60.

[4] Based on ongoing work with Chronicling America, National Digital Newspaper Archive of Mexico, British Library 19th Century Newspapers, Times Digital Archive, Delpher, Europeana, National Library of Finland, Trove and Papers Past.

[5] See Gérard Genette, Paratexts: Thresholds of Interpretation (Cambridge: Cambridge University Press, 2009), and Drew Baker, Anna Bentkowska-Kafel, and Hugh Denard eds. Paradata and Transparency in Virtual Heritage (Farnham: Ashgate, 2012).

[6] METS/ALTO is the most common choice; ALTO (see https://www.loc.gov/standards/alto/) was developed by the METAe to be used alongside METS, and provides layout information. METS provides the structure of the newspaper (see http://www.loc.gov/standards/mets/). Other structures include MPEG-21/DIDL (see ISO/IEC, "Information Technology - Multimedia framework [MPEG-21] - Part 2: Digital Item Declaration", ISO/IEC 21000-2:2005 [October 2005]), a METS/MODS combination (see Morgan Cundiff, "Library of Congress: Using and to Create XML Standards-based Digital Library Applications" [2007] https://www.loc.gov/standards/mods//presentations/litaprogram-an2007.html), METS/PREMIS (see https://www.loc.gov/standards/premis/premis-mets.html), Jisc's own adapted Dublin Core (see http://dublincore.org/documents/dcmi-terms/), Gale legacy XML, the XML used by Gale on their text-mining drives, JSON (see https://www.json.org/), and two different APIs (Trove and Digital New Zealand).

[7] See Patricia Harpring, Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works Online Edition (Los Angeles: Getty Publications, 2010): http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/index.html.

HANDLE: Get a Grip on MALLET

David Lawrence Shepard
shepard.david@gmail.com
UCLA, United States of America

Keywords: topic modeling, text mining, MALLET
Topics: Global, English, Contemporary, text mining and analysis, Literary studies

Large collections of digital text from various sources are now available to researchers and the public, and topic modeling is one of the most common approaches to mining that text. However, existing topic modeling tools are often inaccessible to less technically-adept users. The most popular tools require either using the command line or writing code1. While a GUI2 exists, researchers have expressed frustrations with its limited feature set. HANDLE (Heuristic Analytical Digital Language Environment) is a better GUI for MALLET3 that offers easier text importing, topic model building, visualization, and data exporting, and was developed at UCLA's Scholarly Innovation Lab. Increasing the accessibility of text mining facilitates research at all levels, from undergraduate to faculty, by making it easier to perform research and communicate it to a wider audience.

One-step Importing

HANDLE removes one of the stumbling blocks for first-time users by simplifying importing text. It features a drag-and-drop interface for importing files so that students do not need to learn the command line tools for MALLET. HANDLE can import plain text and

Word documents. Second, a window shows the text as MALLET will understand it: punctuation is removed, and stopwords are replaced with struck-through text. Third, the defaults preserve accent marks and non-Latin characters; more advanced users can adjust these settings. Fourth, stopword settings can be adjusted: a user can drag-and-drop stopword list files into the program, and new stopwords can be added through a contextual menu.

Convenient Experimentation

Most users have to try a few different topic models with different settings before they find one that works, but MALLET only accepts input from the command line, leaving users with the tedious task of retyping commands to change one setting. HANDLE addresses this issue by providing a graphical interface for creating topic models, reducing the number of opportunities for errors. HANDLE also shows a list of all the topic models a user has created and remembers the settings used for each, making comparisons and documentation easier. HANDLE also allows a user to save a project (a collection of documents and topic models) and come back to it later.

Easy Visualization

HANDLE has integrated visualizations. For each topic model, an LDAViz4-inspired interface can be used to explore topic similarity and the top words in each topic. Another interface displays the topics that make up a specific document, and their proportions. These images can be exported for articles or papers.

Export Data into Familiar Tools

Because no tool can anticipate all use cases, HANDLE can export its data in formats that are easy to manipulate. Summaries of all topics within all documents, called "document topic reports" in MALLET, can be exported as spreadsheets. Topic model files can be exported for use with command-line MALLET, or for other computers running HANDLE. HANDLE allows users to focus on the data rather than getting the tool to run, and to work with their data in environments more familiar to them.

HANDLE will be released as an open-source project on GitHub by summer 2020.

Works Cited

Grün, Bettina, and Kurt Hornik, 'Topicmodels?: An R Package for Fitting Topic Models', Journal of Statistical Software, 40.13 (2011)

Jonathan Scott Enderle, Arun Balagopalan, Xiaojing Li, and David Newman, Senderle/Topic-Modeling-Tool: First Stable Release (Zenodo, 2017)

McCallum, Andrew Katchites, MALLET: A Machine Learning for Language Toolkit. (Amherst, MA: UMass Amherst, 2002)

?eh??ek, Radim, and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (Valletta, Malta: ELRA, 2010), pp. 45-50

Sievert, Carson, and Kenny Shirley, LDAvis: Interactive Visualization of Topic Models, version 0.3.2, 2015 [accessed 9 October 2019]

1For example, see Bettina Grün and Kurt Hornik, 'Topicmodels?: An R Package for Fitting Topic Models', Journal of Statistical Software, 40.13 (2011) . and Radim ?eh??ek and Petr Sojka, 'Software Framework for Topic Modelling with Large Corpora', in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (Valletta, Malta: ELRA, 2010), pp. 45-50.

2Jonathan Scott Enderle and others, Senderle/Topic-Modeling-Tool: First Stable Release (Zenodo, 2017) .

3Andrew Katchites McCallum, MALLET: A Machine Learning for Language Toolkit. (Amherst, MA: UMass Amherst, 2002) .

4Carson Sievert and Kenny Shirley, LDAvis: Interactive Visualization of Topic Models, version 0.3.2, 2015 [accessed 9 October 2019].

Time Information System, HuTime — A Visualization and Analysis Tool for Chronological Information of Humanities

Tatsuki Sekino
sekino@nichibun.ac.jp
International Research Center for Japanese Studies, Japan

Keywords: timeline, chronology, web application, calendar
Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, software development, systems, analysis and methods, spatial & spatio-temporal analysis, modeling and visualization, Computer science, Informatics

Time Information System, HuTime - A Visualization and Analysis Tool for Chronological Information of Humanities

Tatsuki Sekino, International Research Center for Japanese Studies

In Digital Humanities, events relating to persons, documents or art works are often reproduced on a computer and compared between them chronologically. Despite these needs, software tools that is available for chronological information from humanities are limited. HuTime is a visualization and analysis tool specialized in the chronological information of humanities.

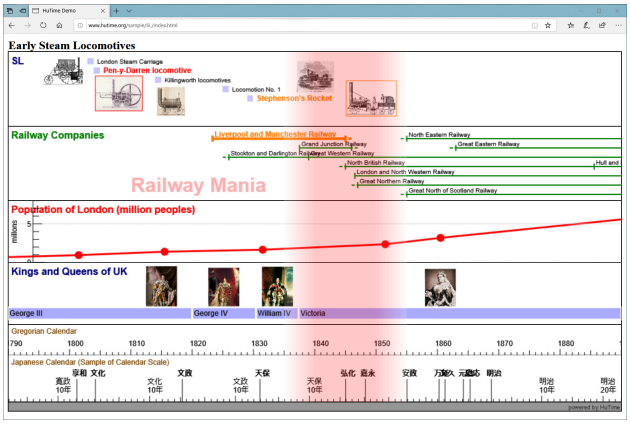


Figure 1: Example of visualizing chronological data1. HuTime has functions to insert images into a timeline and to decorate fonts.

Purpose of HuTime

The time information system, HuTime, aims at providing an environment that includes functions and operations equivalent to geographic information systems for analyzing chronological information (Sekino 2009). HuTime is provided as a JavaScript library2, and it can be accessed using a web browser installed on a personal computer or a smart device such as a tablet (Fig. 1).

Various visualization tools have been developed (e.g., Aigner et al. 2005 and Havre et al. 2000). Additionally, there are other tools used to display a timeline on a web browser (e.g., SIMILE Timeline3, Timeglider4 and TimelineJS5). However, most of these tools display a timeline using data created according to a strict regulation; therefore, processing diversified and complicated data of humanities is difficult. In contrast, HuTime targets humanities and has various functions for data from humanities ("Hu" in the name of "HuTime" originates from the humanities).

Main Features for Humanities

(processing diversified time representations)

As calendar systems differ between countries and historical periods, researchers need to standardize dates in data of humanities to compare them chronologically. Additionally, time periods in humanities are often provided as approximate representations (e.g., "early 17th century") or vague historical periods whose beginning and end points are uncertain (e.g., "Renaissance"). HuTime can process such difficult time representations by applying HuTime Calendar Linked Open Data (Sekino 2017) and a theory for uncertain time (Sekino 2019, 2020), and display these data on a common temporal axis based on the time representations.

(visualizing various contents of the Humanities)

HuTime can chronologically display numerical and image data as well as text data. This function enables researchers to chronologically examine text data, associating it with statistical data such as population dynamics or with photos indicating temporal changes in historical materials such as a cultural asset.

HuTime can directly process data from different web sites on the internet (Fig. 2). Recently, various kinds of open data from humanities are increasing on the internet; therefore, HuTime can utilize them effectively.

Application

HuTime which provides chronological functions is available as a library for various web applications (Fig. 2). For example, when HuTime is embedded in a database application as a user interface, the application displays a retrieval result on the embedded HuTime (e.g., Ichino et al. 2018).

HuTime is an open source software whose source code is available on GitHub6, and is expected to be a platform for chronological information equivalent to Google Maps, which provides a platform for geographic information.

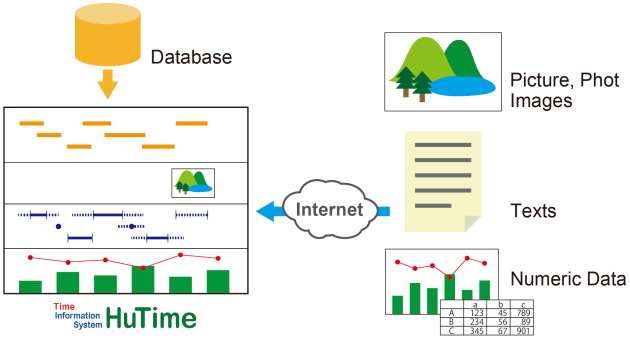


Figure 2: HuTime can visualize a retrieval result of a database, associating it with various resources on the internet.

Bibliography

1. Aigner, W., Miksch, S., Thurnher, B. and Biffl, S. (2005) PlanningLines: novel glyphs for representing temporal uncertainties and their

evaluation. Proceedings od Ninth International Conference on Information Visualisation, pp. 457-463.

2. Havre, S., Hetzler, B. and Nowell, L. (2000) ThemeRiver: visualizing theme changes over time. Proceedings of IEEE Symposium on Information Visualization 2000, pp. 115-123.

3. Ichino, M., Hashimoto, Y., Hirano, J., Masuda, K. and Kitamoto, A. (2018) A pathfinder system for Historical Situation Records using eyewitness information. Proceedings of the Computers and the Humanities Symposium pp.343-350.

4. Sekino, T. (2009) Tools to Realize Spatiotemporal Analysis in the Humanities. Proceedings of GIS in the Humanities and Social Sciences International Conference pp.151-159.

5. Sekino, T. (2017) Basic linked data resource for temporal information. Proceedings of the 2017 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC) pp.76-82.

6. Sekino, T. (2019) Using Uncertain Time Intervals in Linked Data. International Journal of Geoinformatics 15(2):15-23.

7. Sekino, T. (2020) Data description and retrieval using periods represented by uncertain time intervals. Journal of Information Processing 28:91-99.

Notes

1. HuTime Demo. http://www.hutime.org/sample/2020/SL/

2. Web HuTime. http://web.hutime.org/

3. SIMILE Timeline. http://www.simile-widgets.org/timeline/

4. Timeglider. https://www.preceden.com/

5. TimelineJS. http://timeline.knightlab.com/

6. HuTime Project in GitHub. https://github.com/hutime/

Teaching Across the Digital Divide: The Challenges of Doing Digital Humanities in the Community College Classroom

Orion Teal
oteal@elcamino.edu
El Camino College Department of History, United States of America

Keywords: digital pedagogy, community colleges, equity
Topics: English, North America, 20th Century, Contemporary, curricular and pedagogical development and analysis, digital access, privacy, and ethics analysis, Chicano/a/x, Latino/a/x studies, History

The community colleges are fertile ground for ongoing efforts to democratize digital humanities, but infrastructural challenges and concerns about digital surveillance pose unique problems for instructors pursuing digital pedagogy in the community college classroom. This is particularly true at campuses that serve economically disadvantaged students with limited access to computers and unreliable internet connections at home. Students from historically underrepresented groups, who make up the bulk of community college students in my region, also face challenges in completing digital projects, including apprehensions about maintaining a public digital footprint due to immigration status. This poster presentation explores how the digital divide, specific concerns of underrepresented students, and technology infrastructure limitations on campus have shaped my attempts at digital pedagogy at a leading community college in the Los Angeles area.

From data to visualisation: Dante's Divine Comedy as a case study. 1

Ginestra Ferraro
ginestra.ferraro@kcl.ac.uk
King's College London, United Kingdom

Keywords: data visualisation, sentiment analysis, spatio-temporal textual data
Topics: Global, English, 5th-14th Century, 15th-17th Century, Contemporary, software development, systems, analysis and methods, spatial & spatio-temporal analysis, modeling and visualization, Computer science, Humanities computing

A journey from Hell to Heaven, investigating the computational opportunities of automating text analysis and producing data visualisations.

This poster presents the results of the exploratory work for a reusable tool to generate data visualisations

based on automatic text analysis.2 Its non-functional requirements respond mainly to flexibility (accept different text inputs) and optimisation (produce rich visualisations with minimal set up). The visual outputs produced by the application have an explorative function in that they aim to:

- offer a different perspective on the text under study;
- highlight patterns and/or outliers (Meirelles 2013);
- drive research in formulating new hypotheses;
- provide support to, or disprove, existing theses.

The current version accounts for modules (i.e. software components) designed around one selected
test case, namely Dante Alighieri's Divine Comedy, but serves as a blueprintfor further modules to be plugged in.

The Italian version of the Commedia (Petrocchi 1966-67) is used to perform text structural analysis and work on the rhyme scheme, while the English translation (Mandelbaum 1980-84) is used for sentiment analysis. The unique way in which Dante wrote his masterpiece, makes the text an interesting dataset to be explored computationally. Structural (spatial and temporal)
textual components lend themselves to be represented graphically, and offer insights into its linguistic content.The visual outputs allows users to interact with both the content and the metadata.

The application performs computational text analysis to produce data visualisations representing the following structural, stylistic and semantic features of the text:

- schematic representation of the poem's structure and rhythm (fig. 1);
- distribution of keywords (fig. 2);
- visual
  representation of the sentiment analysis (fig. 3).

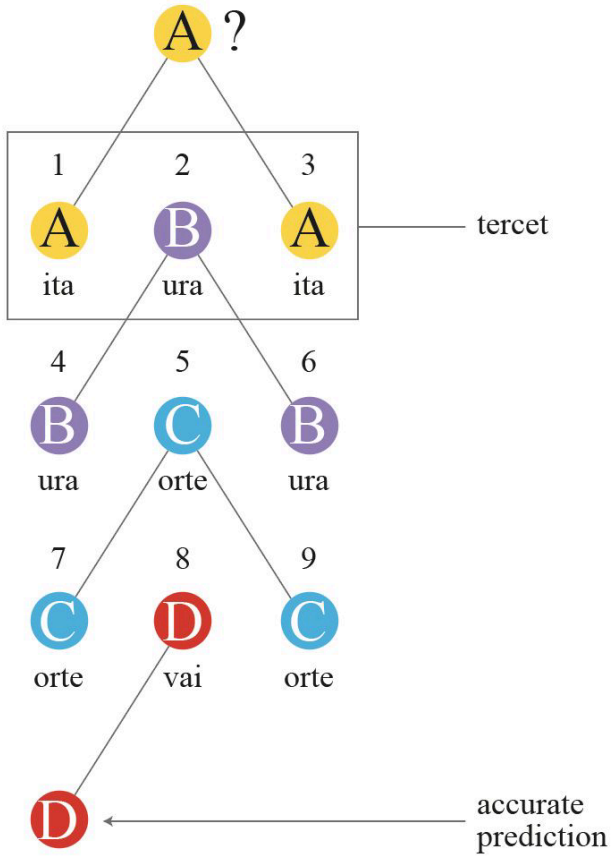


Figure 1 An example of the schematic representation of the poem's structure: rhythm imposed by tercets and rhyme prediction.

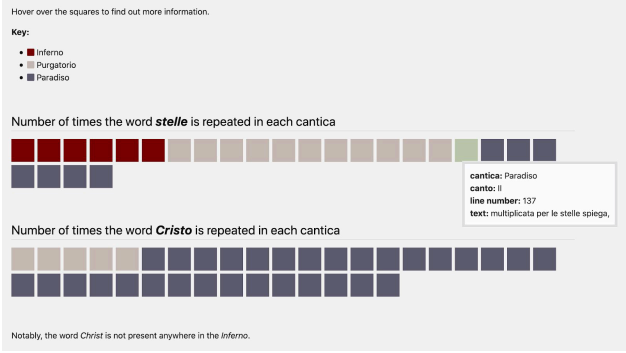


Figure 2 Words like Cristo (Christ) and stelle (stars) are distributed unevenly across the three cantiche: the word "Christ" never appears in the Inferno, while it's widely used in the Paradiso. One square per line.

Figure 3 Sentiment analysis visualisation of the three cantiche. Red is negative, blue is positive and the opacity indicates how close to the polarity (-1, 1) the sentiment is. One square per line.

The application has been developed modularly (Martin and Martin 2006), following the separation of concerns design principle (Dijkstra 1982) to allow for flexibility and scalability.The computational aspect of the project is implemented in Python, a flexible programming language that supports object-oriented programming and functional paradigms.The visualisations are produced with the support of d3.js library, "a JavaScript library for manipulating documents based on data" (Bostock D3.js). The application exploits HTML5 and SVG specifications to allow for greater interaction and portability.Natural language processing (NLP) and machine learning techniques have been applied to process and transform the data. The Naive Bayes Classifier (Perkins 2010) technique has been chosen due to its performance and simple implementation.A training dataset has been manually created collecting random subsets of text from other authors close in language and time, and further work from Dante himself:
Ludovico Ariosto, Orlando furioso (Wikisource contributors 2012)Dante Alighieri, Convivio (Gallarino)Giovanni Boccaccio, Decamerone (Wikisource contributors 2017)

The poster illustrates the workflow from input to output, displaying a diagram of the process.

The poster demonstrates achievements of this proof of concept and development ideas for the future. The main success lies in its modular development (fig. 4), making it amenable to further development3 (algorithm refinements, visualisation workflows, stylometric analysis). More languages and different text structures will be integrated and a wider range of output visualisations offered, while making use of the same core functionalities for ingesting and processing data.

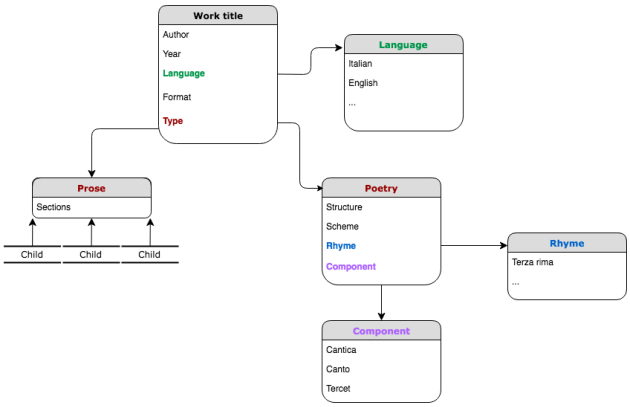


Figure 4 The data model of the application, illustrating the separation of concerns andthe potential for extensibility.

Footnotes1 Github repository: https://github.com/ginestra/dante-visualised2 Existing tools such as Voyant Tools (https://voyant-tools.org/), HuViz (https://cwrc.ca/Documentation/HuViz/HuViz_Concept.html), iteal (http://iteal.vizcovery.org/) are either collating a wide range of existing tools, or targeting very specific cases. The aim of the tool presented here is to provide a lightweight, but solid application core that is extensible in the hands of the user, but doesn't lose its simplicity in use at set up.3 The project is supported by King's Digital Lab (KDL) 10% time allocated for personal research, aiming atfurther funding for future development in order to guarantee the extension and long term sustainability ofthe application.

List of references

Meirelles I., "Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations", Rockport Publishers, 2013, p. 189.Petrocchi, G. (1966-67). "Dante Alighieri, La Commedia secondo l'antica vulgata, a cura di Giorgio Petrocchi", Mondadori, Milano.Mandelbaum, A. (1980-84). "Divine Comedy of Dante Alighieri", University of California Press.Martin R. C., Martin M., "Agile Principles, Patterns, And Practices in C#", Prentice Hall, 2006, ch. 8.Dijkstra, E. W. (1982). "On the role of scientific thought". Selected writings on Computing: A Personal Perspective . New York, NY, USA: Springer-Verlag. pp. 60-66. ISBN 0-387-90652-5.Bostock, M., D3.js < https://d3js.org/ > [Accessed 12 September 2019]Perkins

J., "Text classifier for sentiment analysis - Naive Bayes Classifier", 10 May 2010, [Accessed 12 September 2019]Wikisource contributors (2012), "Orlando Furioso/Canto V", Wikisource, [accessed 12 September 2019]Gallarino, M., "Convivio" [accessed 12 September 2019]Wikisource contributors (2017), "The Decameron (Rigg)", Wikisource, [accessed 27 November 2018]

Visualizing Objects, Places, and Spaces: A Digital Project Handbook

Beth Fischer
blf1@williams.edu
Williams College Museum of Art

Hannah L. Jacobs
hannah.jacobs@duke.edu
Duke University / University of North Carolina-Chapel Hill

Keywords: planning, visual, projects, resources, methods
Topics: Comparative (2 or more geographical areas), Global, English, North America, 5th-14th Century, 19th Century, Contemporary, project design, organization, management, sustainable procedures, systems, and methods, Art history, Education/ pedagogy

Digital humanities are rich with publications, workshops, guides, and resource lists that introduce the field's concepts and methods. As digital humanities have grown, trainings and materials for experienced practitioners have become prominent. Yet, as the authors have observed through their involvement with the National Humanities Center's Summer Institute on Objects, Places, and the Digital Humanities, needs remain for resources that introduce anyone interested in creating a dh project to core methodological considerations and knowledge necessary to productively choose platforms and approaches that fit their research goals. This gap is especially apparent in fields dealing with visual materials and physical spaces, such as digital art history, where infrastructure for visual technologies varies. Visualizing Objects, Places, and Spaces: A Digital Project Handbook (https://handbook.pubpub.org) is designed to address that gap. This poster outlines the Handbook's structure, reflects on the project's own development process, and shares next steps.

"Women's Writing in the Eighteenth Century Evaluating 'Representative' Corpora"

Lawrence Isaac Evalyn
lawrenceevalyn@gmail.com
University of Toronto, Canada

Keywords: text mining, corpus building, eighteenth century, literary canons, gender
Topics: Europe, English, 18th Century, database creation, management, and analysis, text mining and analysis, Feminist studies, Literary studies

Despite the crucial importance of corpus-building to the interpretation of text-mining research, it is often extremely difficult to know what is in a corpus. Even large institutional resources used by many scholars provide little context for their choices of what to include or exclude. These hidden choices are particularly problematic when historical selection factors might have led to the creation of corpora which re-create social inequalities. I examine six corpora which are used as the basis of most eighteenth century distant reading. I manually evaluate each corpus's holdings for a very narrow selection of texts, works published in England 1789-99, to answer a series of bibliographical questions, including: how many titles are by men, by women, or unsigned? What broad categories of writing are represented - novels, plays, poetry, pamphlets, songs, sermons, ephemera, others? Analyzing the differences, I ask: do the most invested-in resources underrepresent women?

Distant Viewing Toolkit Software for the Analysis of Visual Culture

Lauren Tilton
ltilton@richmond.edu
University of Richmond, United States of America

Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Keywords: media studies, computational social science, time-based media, visual culture, computer vision
Topics: Global, English, 20th Century, Contemporary, image processing and analysis, music and sound digitization, encoding, and analysis, Computer science, Media studies

The poster will introduce the Distant Viewing Toolkit (DVT), a Python package for the computational analysis of visual culture. The toolkit addresses the challenges of working with moving images through the automated extraction and visualization of metadata summarizing the content (e.g., people/actors, dialogue, scenes, objects) and style (e.g., shot angle, shot length, lighting, framing, sound) of time-based media. DVT is optimized for two purposes: scholarly inquiry of visual culture from the humanities and social sciences, and search and discovery of collections within libraries, archives, and museums.

THOTH – Transcribing historical objects with tabulated handwriting

Oliver Dunn
od226@cam.ac.uk
Cambridge University, United Kingdom

Alexis Litvine
adl38@cam.ac.uk
Cambridge University, United Kingdom

Keywords: machine learning, computer vision, historical data
Topics: Global, English, 15th-17th Century, 18th Century, 19th Century, artificial intelligence and machine learning, database creation, management, and analysis, Computer science, History

We present a new computer vision and machine learning tool for extracting data from historical sources. This tool has application in the humanities and the social sciences, public archives and libraries. It is being developed as a collaboration between Cambridge-based researchers in history and computer sciences. It develops on existing machine learning tech for manuscript recognition to machine read historical table structures and to more rapidly extract these into modern relational databases.

Re-interpreting the Paintings and Poems of the Lotus Sutra by a Co-occurrence Network: The Religion and Life of Aristocratic during the Japanese Medieval Ages

Aiko Aida
nispanna@yahoo.co.jp
Japan society for the Promotion of Science

Keywords: Buddhist art, Japanese paintings, Poetry
Topics: Asia, English, 5th-14th Century, attribution studies and stylometric analysis, natural language processing, Art history, Theology and religious studies

This presentation intends to explore the correlation between Waka literature and the paintings of the whole Lotus Sutra in Japan from the eleventh to fourteenth centuries, as derived from a correspondence analysis between the two, using extracted vocabularies of annotation and visualization by applying a dimension reduction. In addition, I mainly analyze the twenty-one imperially commissioned poetry anthologies for the literature of the Lotus Sutra, and for paintings, I focus on the frontispieces of the Lotus Sutras written in gold and silver inks on indigo-dyed papers.

In particular, I attempt to reinterpret the visualization of dimension reduction by using a new method such as t-SNE (t-distributed Stochastic Neighbor Embedding? in addition to conventional PCA (principal component analysis) to emphasize the differences instead of similarities.

This may further reveal the literary peculiarities of the rhetoric, the nature of the "narrative", or the visual arts-specific rhetoric such as "decoration" and newer contexts.

Increasing Sacred Site Accessibility for Indigenous Elders through Virtualization: Demonstrating Community Engagements at the Mesa Prieta Petroglyph Project

Chester Ronald Liwosz
chester@mesaprietapetroglyphs.org
Mesa Prieta Petroglyph Project, United States of America

Keywords: virtual heritage managment, direct action, youth engagement
Topics: English, North America, 15th-17th Century, Contemporary, digitization (2D & 3D), virtual and augmented reality creation, systems, and analysis, Anthropology, Archaeology

The community-based not-for-profit Mesa Prieta Petroglyph Project (MPPP) has expressly sought to engage with local and indigenous residents, especially Tewa whose ancestors produced the images largely believed to date to the Rio Grande Classic Period (ca. AD 1350 to 1600). Two challenges surrounding contemporary engagement involve reaching both youth and elders. Presently, MPPP is piloting ways the technologies we deploy in our research and documentation can be used to address these challenges. We share how we both engage youth in the region's cultural history, and enable limited-mobility elders' engagement through 3D multimedia experiences, all using new smart technologies. Both endeavors integrating contemporary technologies with our indigenous outreach help democratize archaeological practice by empowering members of Native America communities to take possession of, and to adapt and retell, their ancestors' stories. Alongside this poster we present demonstrations of these technologies, offering conference attendees firsthand experience with our new digital outreach efforts.

Statistical Traces of Censorship in Russian Political Thought

Mark Schwindt
mark.schwindt@rub.de
Ruhr-University Bochum, Germany

Keywords: censorship, Russia, conceptual history, political thought, intelligentsia
Topics: Europe, English, 19th Century, 20th Century, bibliographic analysis, text mining and analysis, Cultural studies, Philosophy

The usage of word counts is still a valuable tool for the non-linear reading of individual works or mid-sized corpora, especially in digital conceptual history. We present first results of a larger project focusing on the history of political concepts in 19th and 20th century Russia, in particular on roughly 300 large essay collections, so-called 'ideynye sborniki'. Our poster discusses word counts and co-occurrences in the table of contents of the manifestos in order to identify the purported topics of discourse in these texts and compares them to the usage of these terms in the texts themselves, i. e. their actual content.

DH, Disciplinarity, and the Republic of Tweets

Jessica Otis
jotis2@gmu.edu
George Mason University

Keywords: Twitter, social network analysis, disciplinarity
Topics: Global, English, North America, Contemporary, network analysis and graphs theory and application, social media analysis and methods, Cultural studies

Since Twitter's creation in 2006, the social media platform has become a high-profile conduit for backchannel conversations and the public dissemination of information during academic conferences and events. While some tweets simply vanish into the internet void, others provoke lively

conversation within the conference community or are retweeted to signal boost them into adjacent or similar communities. The networks formed by these conference tweets-a 21st-century Republic of Letters written 140 or 280 characters at a time-can be put in conversation with one another to reveal interdisciplinary patterns of associations among digital humanists.

Better preserving the European Silk Heritage with the SILKNOW ontology. A CRM extension for modelizing the production process of silk artefacts

Marie Puren
marie.puren@cnrs.fr
Laboratoire de recherche historique Rhone-Alpes / CNRS, France

Pierre Vernus
pierre.vernus@ish-lyon.cnrs.fr
Laboratoire de recherche historique Rhone-Alpes / Université Lumière Lyon 2, France

Keywords: CIDOC CRM, history of technics, textile, linked data
Topics: Europe, English, 15th-17th Century, 18th Century, 19th Century, data modeling, data, object, and artefact preservation, History

Silk played an important role in European history, mostly along the Western Silk Road's network of production and market centres. Although many European specialized museums are devoted to its preservation, they usually lack size and resources to establish networks or connections with other collections. The H2020 SILKNOW project (Silk heritage in the Knowledge Society: from punched card to Big Data, Deep Learning and visual/tangible simulations) aims to produce an intelligent computational system in order to improve our understanding of European silk heritage

The SILKNOW platform

The poster introduces the scope and goals of the H2020 SILKNOW project and presents some of the functionalities offered by the online platform. The platform has been designed to freely and easily give access to a wide variety of aggregated datasets, describing silk-related artefacts produced in Europe between the 15th and the 19th century, and preserved in Cultural Heritage Institutions (CHIs) around the world.

The definition of the SILKNOW ontology

This computational system is modeled and trained thanks to datasets crawled by SILKNOW from online databases of CHIs, or provided by them. To aggregate these various datasets, it is necessary to harmonize them by designing and implementing a unique and complete data model. This data model is based on the CIDOC Conceptual Reference Model (CIDOC-CRM). The classes and properties selected for the SILKNOW data model are publicly accessible and documented via OntoMe, an ontology management system, developed by the LARHRA research center (http://ontome.dataforhistory.org/). After evaluating the goodness of the ontology by providing mapping rules between CHI's metadata records and the SILKNOW ontology, we observed that all fields can be represented by using existing classes and properties from the SILKNOW ontology so far.

A CRM extension to describe the production process of silk artefacts

CIDOC-CRM is also a core ontology with more specialist extensions. In other words, it is possible to add new subclasses and sub-properties to express more specific relationships and properties, without modifying the basic structure of the model.

The complex modeling of the semantics included in data about the creative and productive process of silk textiles cannot accurately be mapped with the SILKNOW ontology. Free-text fields are generally used by CHIs to analyse the structure and the decoration of the fabrics, or to present the historical context of their production or their use. This first mapping aimed at storing these metadata "as they are".

The more complex modeling of the semantics included in data about the creative and productive process of silk textiles requires elaborating new classes and properties. There is yet no CRM extension for dealing with the production of textile artefacts, something similar to FRBRoo, for the creation, production and expression process in literature and the performing arts. We present the CRM extension we are currently elaborating for this purpose. A complete overview of these new classes and properties is publicly available via Ontome : http://ontome.dataforhistory.org/profile/10

References

CIDOC-CRM Homepage, http://www.cidoc-crm.org/. Last accessed 07 October 2019.

FRBRoo - Functional Requirements for Bibliographic Records Model, http://www.cidoc-crm.org/frbroo/home-0. Last accessed 07 October 2019.

Le Boeuf, P., Doerr, M., Ore, C., Stead, S.: Definition of the CIDOC Conceptual Reference Model, Produced by the ICOM/CIDOC Documentation Standards Group, Continued by the CIDOC CRM Special Interest Group Version 6.2.1, October 2015, http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.1.pdf. Last accessed 07 October 2019.

Ontology Management Environment (OntoME), http://ontome.dataforhistory.org/. Last accessed 14 October 2019

SILKNOW Homepage, http://silknow.org/. Last accessed 07 October 2019.

By the People at the Library of Congress

Victoria Anne Van Hyning
vvanhyning@loc.gov
Library of Congress, United States of America

Keywords: crowdsourcing, open data, public humanities, libraries, cultural heritage
Topics: Global, English, 15th-17th Century, 19th Century, 20th Century, crowdsourcing, public humanities collaborations and methods, Library & information science

In October 2018 the Library of Congress launched By the People (https://crowd.loc.gov/), a text transcription and tagging crowdsourcing project powered by a new platform called Concordia. Concordia was developed at the Library to accommodate the particular needs of the Library's infrastructure and legacy data management systems. This poster will focus on the data lifecyle of By the People transcriptions, and offer a possible model for practitioners at other institutions.

Analysis and Categorisation of Research Software in the Digital Humanities

Lukas Harzenetter
lukas.harzenetter@iaas.uni-stuttgart.de
Institute of Architecture of Application Systems, University of Stuttgart, Germany

Johanna Barzen
Institute of Architecture of Application Systems, University of Stuttgart, Germany

Frank Leymann
Institute of Architecture of Application Systems, University of Stuttgart, Germany

Brigitte Mathiak
Data Center for the Humanities, University of Cologne, Germany

Philip Schildkamp
Data Center for the Humanities, University of Cologne, Germany

Claes Neuefeind
Cologne Center for eHumanities, University of Cologne, Germany

Uwe Breitenbücher
University of Stuttgart, Germany

Keywords: Living Systems, Research Applications, Software Stacks, Sustainability, TOSCA
Topics: Global, English, Contemporary, digital libraries creation, management, and analysis, systems and information architecture and usability, Humanities computing, Informatics

Abstract

In the Digital Humanities (DH) research software is a major output of the scientific process besides primary research data. However, long-term sustainability and preservation of living systems, in contrast to long-term archiving (LTA) of research data, is still a non-trivial, often institutionally segmented enterprise (Smithies et al., 2019). To counter this, the SustainLife project, running at the Institute of Architecture of Application Systems (IAAS, University of Stuttgart) and the Data Center for the Humanities (DCH, University of Cologne), uses the Topology and Orchestration Specification for Cloud Applications (TOSCA) standard (OASIS, 2013 and 2019), and the open source ecosystem OpenTOSCA

(Breitenbücher et al. 2016). To realize our vision of a sustainable long-term DH research software archive, we are building a repository of software components modeled in TOSCA, i.e., Node Types, and use them to create application blueprints, i.e., Topology Templates in TOSCA, describing the structure of an application.

To identify the most used software component types within the DH, we approach the vast field of DH research software from multiple perspectives (Neuefeind et al., 2018 and 2019): (1) We investigate multiple use cases in depth and extract the components used in the respective applications, starting from the operating system (OS) level up to the user interface (UI) layer. These components are then modelled as reusable Node Types in TOSCA which can be utilized to describe the respective application in a TOSCA Topology Template. (2) Qualitative case studies aim at extracting all employed components and technologies of selected applications and (3) quantitative surveys, targeted at DH research software experts (i.e. researchers), are performed to collect community practices and demands. The results from both the case studies and the surveys are then used to identify key components and derive application stacks, i.e., Node Types and Topology Templates.

Our poster presents a "Categorisation of Research Software in the Digital Humanities" to spawn discourse about the perspectives of DH research software in contrast to research data (Sahle and Kronenwett, 2013) and how this vast landscape can be indexed in a way that researchers may find the most appropriate set of component types or a suitable application blueprint, when searching for their own software. In this context, we investigate the use of categorial keywords which, in combination with technological identifiers, e.g., programming languages and database types, can be used to provide a means for categorizing research software in the DH. Therefore, our poster will present selected DH research software as a basis for discussion. Hereby, we introduce component types and application stacks from our repository, e.g., an Apache web server or the application stack of a DH research software respectively. Moreover, a set of technological identifiers and categorial keywords, such as digital editions, virtual research environments, and interactive visualisations (Wuttke et al., 2016), associated with these projects will be presented.

Acknowledgements

This Poster is partially funded by the DFG-LIS project "SustainLife" (GEPRIS 379522012).

References

Breitenbücher, U. and Endres, C. and Képes, K. and Kopp, O. and Leymann, F. and Wagner, S. and Wettinger, J. and Zimmermann, M. (2016). The OpenTOSCA Ecosystem. Concept & Tools. In: European Space project on Smart Systems, Big Data, Future Internet - Towards Serving the Grand Societal Challenges - Volume 1: EPS Rome 2016. SciTePress, pp. 112-130.

Neuefeind, C. and Schildkamp, P. and Mathiak, B. and Harzenetter, L. and Barzen, J. and Breitenbücher, U. and Leymann, F. (2018). Technologienutzung im Kontext Digitaler Editionen. Eine Landschaftsvermessung. In: Book of abstracts of the 6th annual conference of the Digital Humanities im deutschsprachigen Raum (DHd 2019), pp. 219-222.

Neuefeind, C. and Schildkamp, P. and Mathiak, B. and Mar?i?, A. and Hentschel, F. and Harzenetter, L. and Breitenbücher, U. and Barzen, J. and Leymann, F. (2019). Sustaining the Musical Competitions Database. A TOSCA-based Approach to Application Preservation in the Digital Humanities. In: Book of Abstracts of the 29th Digital Humanities Conference (DH 2019), https://dev.clariah.nl/files/dh2019/boa/0574.html (retrieved: 2019-09-10).

OASIS (2013). Topology and Orchestration Specification for Cloud Applications Version 1.0, http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html (retrieved: 2019-09-10).

OASIS (2019). TOSCA Simple Profile in YAML Version 1.2, http://docs.oasis-open.org/tosca/TOSCA-Simple-Profile-YAML/v1.2/TOSCA-Simple-Profile-YAML-v1.2.html (retrieved: 2019-09-10).

Sahle, P. and Kronenwett, S. (2013). Jenseits der Daten. Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner Data Center for the Humanities. In: LIBREAS. Library Ideas #23, pp. 76-96.

Smithies, J. and Sichani, A. M. and Westling, C. and Mellen, P. and Ciula, A. (2019). Managing 100 Digital Humanities Projects. Digital Scholarship & Archiving in King's Digital Lab. In: Digital Humanities Quarterly, http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html (retrieved: 2019-09)

Wuttke, U. and Engelhardt, C. and Buddenbohm S. (2016). Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum. In: Zeitschrift für digitale Geisteswissenschaften,? pp. 1-12.

No comment: Addressing comment sections in web analysis

Paul Guille-Escuret
paul.guille.escuret@gmail.com
CNRS / Université Paris Sorbonne (GEMASS) Aix-Marseille (VITROME)

Florian Cafiero
florian.cafiero@gmail.com
CNRS / Université Paris Sorbonne (GEMASS)

Jeremy Ward
CNRS / Université Paris Sorbonne (GEMASS) Aix-Marseille (VITROME)

Keywords: internet studies, network analysis
Topics: Global, Europe, English, 20th Century, Contemporary, network analysis and graphs theory and application, social media analysis and methods, Humanities computing, Sociology

We provide an R-based method for extracting the commentary sections of a website - the contents of which can bias a corpus' analysis, but can be interesting to study per se.

Studying corpora of websites, through methods such as topic modeling or hyperlink analyses, is an increasingly adopted approach in the humanities (e.g. Severo et al., 2018, Romele et al. 2016, Berthelot et al. 2016), information science (e.g. Bounegru et al., 2017) or social science (Marres, 2015, Froio, 2018). Yet, one part of their content is very often neglected: the comments section.

The biases induced from leaving the comments section

Especially when studying a corpus of websites focusing on controversial topics, commentary sections can induce many biases in the analyses. Comments can express a point of view radically different from the page itself. Hyperlinks present in the comments can point to contents that the owner of the website does not endorse, which can distort any network analysis. The vocabulary used in the comments can also bias content analyses such as topic modelling. It is thus key to eliminate these comments, or to keep them for a separate analysis. We exemplify this through a case study.

Separating the comments from the page: a tedious task

Removing or extracting the commentary sections from a set of websites is in fact a tedious task, thus rarely performed. Many languages can be used to encode the page: HTML 4.0 or 5.0, XHTML, Ajax, Ruby on Rails etc. Some standards obviously exist, for instance for blog platforms, but they are not widely adopted. And unexpected means to open a commentary section (e.g. considering the commentary sections as a subpart of a forum) can frequently occur.

Aiming at exhaustivity: a necessity

Focusing only on the easily retrievable commentary sections would induce important biases. The way the commentary section is encoded is in itself a socially-induced phenomenon, demonstrating the user's literacy in web programming, or his financial means. Excluding very poorly encoded pages, or virtuoso contents written by expert programmers, could thus translate into excluding specific groups from any further analysis.

A method for extracting comments

The method we propose is not fully automated, and requires a direct identification of patterns delimiting comments sections and comments themselves in the code. Some patterns are relevant for many websites while others need to be carefully designed for a single use. We then provide an implementation with R of a code which carries out the rest of the procedure: after automated quality checks and potential improvements, links and contents coming from comments are subtracted, and the comment-free pages can be analysed. Comment sections themselves can be extracted for a separate analysis.

REFERENCES

BERTHELOT, Marie-Aimée, SEVERO, Marta, et KERGOSIEN, Eric. Cartographier les acteurs d'un territoire: une approche appliquée au patrimoine industriel textile du Nord-Pas-de-Calais, 2016.

BOUNEGRU, Liliana, VENTURINI, Tommaso, GRAY, Jonathan, et al. Narrating Networks: Exploring the affordances of networks as storytelling devices in journalism. Digital Journalism, 2017, vol. 5, no 6, p. 699-730.

FROIO, Caterina. Race, religion, or culture? Framing Islam between racism and neo-racism in the online network of the french far right. Perspectives on Politics, 2018, vol. 16, no 3, p. 696-709.

MARRES, Noortje. Why map issues? On controversy analysis as a digital method. Science, Technology, & Human Values, 2015, vol. 40, no 5, p. 655-686.

ROMELE, Alberto et SEVERO, Marta. From Philosopher to Network. Using Digital Traces for Understanding Paul Ricoeur's Legacy. Azimuth. Philosophical Coordinates in Modern and Contemporary Age, 2016, vol. 6, no 6.

SEVERO, Marta and VENTURINI, Tommaso. Intangible cultural heritage webs: Comparing national networks with digital methods. New Media & Society, 2016, vol. 18, no 8, p. 1616-1635.

Colaboración recíproca - ¿Cómo los proyectos de humanidades digitales pueden retribuir a las instituciones culturales?

Ernesto Priani Saisó
epriani@gmail.com
Universidad Nacional Autónoma de México (UNAM), Mexico

Isabel Galina Russell
igalina@unam.mx
Universidad Nacional Autónoma de México (UNAM), Mexico

Keywords: newspaper collections, open data, open access, data mining, national library
Topics: Spanish, North America, Contemporary, digital libraries creation, management, and analysis, meta-criticism (reflections on digital humanities and humanities computing), Library & information science

¿Pueden los proyectos de humanidades digitales (HD) retribuir a las instituciones culturales que les proporcionan los materiales que estudian? En particular, en América Latina, ¿los proyectos de humanidades digitales pueden ser un medio para enriquecer los acervos digitales, y cumplir con ello un propósito social? ¿De qué forma un proyecto internacional tiene impactos nacionales?

NEH Funding for Digital Projects

Sheila A Brennan
sbrennan@neh.gov
National Endowment for the Humanities, United States of America

Keywords: funding
Topics: English, North America, Contemporary, project design, organization, management, public humanities collaborations and methods, Education/ pedagogy, Humanities computing

It can be overwhelming for US-based digital humanities scholars to navigate through the National Endowment for the Humanities's (NEH) extensive list of funding opportunities to find the right program to fund their digital projects. Since the Office of Digital Humanities (ODH) began funding digital projects in 2008, complementary grant programs have expanded while others emerged across the Endowment to support digital work. ODH remains the home for funding experimental, innovative digital methods and infrastructure projects in humanities research, teaching and learning, public digital humanities, and scholarly communications. Yet, it is only one of many divisions with grant programs that might fund your next project. The poster will be designed with questions prompting visitors to address the goals, methods, audiences, and final products of their digital projects to guide them to finding an appropriate funding opportunity. Brennan will advise and discuss opportunities available for individuals and organizations.

Extending the Utility of the HTRC Extracted Features Dataset Through Linked Data

Jacob Jett
jjett2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Boris Capitanu
University of Illinois at Urbana-Champaign, United States of America

Deren Kudeki
University of Illinois at Urbana-Champaign, United States of America

Timothy W. Cole
University of Illinois at Urbana-Champaign, United States of America

J. Stephen Downie
University of Illinois at Urbana-Champaign, United States of America

Keywords: HathiTrust, datasets, non-consumptive analysis, extracted features
Topics: Global, English, Contemporary, artificial intelligence and machine learning, data publishing projects, systems, and methods, Computer science, Library & information science

A poster describing the latest version of the Extracted Features dataset derived from the HathiTrust Digital Library's 17+ million volume corpus. This version employs Linked Data standards to both, make the dataset more accessible and to incorporate richer metadata describing the volumes from which the data was derived. The dataset is arranged by the volumes and the data (tokens, part of speech tags, language tags, line counts, etc.) is directly associated with the metadata describing the volume in the form individual JSON-LD documents. The EF dataset provides a ready means of interacting with volumes whose intellectual content remains under copyright and allows a variety of analytics, such as visualizing word usage over time, to be carried out on data that would not otherwise be accessible.

Automatic Extraction of Poetry from Digitally Scanned Books

John Foley
jjfoley@smith.edu
Smith College, United States of America

Keywords: Poetry Extraction, Information Retrieval, Machine Learning
Topics: English, North America, 15th-17th Century, 18th Century, 19th Century, digital libraries creation, management, and analysis, information retrieval and querying algorithms and methods, Book and print history, Computer science

We present an automatic, learned model for the extraction of poetry from digitally scanned books. This abstract highlights our recent work on poetry identification from Internet Archive books and the public resources (code, data and models) that exist as a result. We hope that this is the beginning of deeper and richer research into poetry in the digital humanities because curating custom collections of poetry should be less expensive.

Poetry in Digital Libraries

Digital libraries have expanded rapidly in quantity and quality of content over the past decade. Out-of-copyright and public domain works are available from the inventing of the printing press all the way to the early twentieth century.

Unfortunately, this explosion in content has not quite connected all the way to different genres: large collections of poetry are not available because they are typically curated manually.

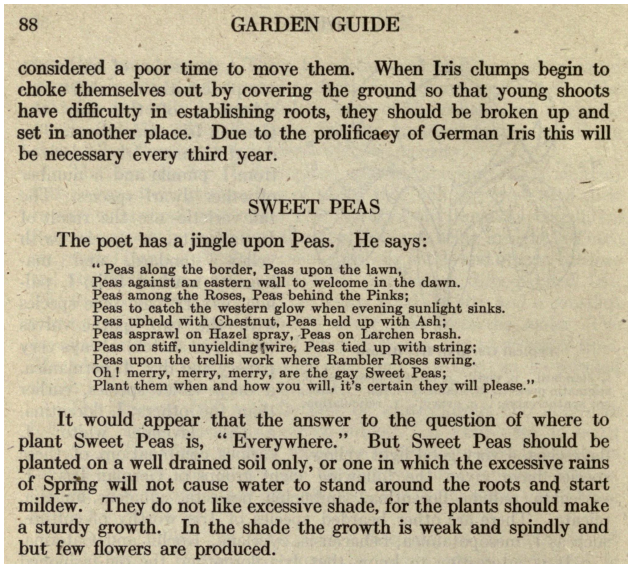
The intersection of poetry and digital methods is actually fairly common and has been studied in a diverse set of languages and cultures e.g., Bangla (Rakshit et al., 2015), Arabic (Ahmed and Trausan-Matu, 2017) and Thai (Promrit and Waijanya, 2017). Features of poetry have also been studied using computational methods, e.g, meter (Hamidi et al., 2009), style (Baumann et al., 2018), authorship and time (Can et al., 2011), emotion (Alsharif et al., 2013; Barros et al., 2013; Kumar and Minz, 2014), and even content (Jamal et al., 2012; Choi et al., 2016; Lou et al., 2015; Kesarwani, 2018). Kaur an Saini's recent work on classifying Punjabi poems into four categories is not a survey, but does provide a table of recent work, language targeted, and features discussed (2017).

However, most of these works use small datasets (10s-100s of poems), because the cost of collecting and curating poetry is so high. There is a lot of poetry available in digital libraries, but it's effectively hidden in those books.

Automatic Extraction of Poetry

Underwood et al. (2013) present a study of genre in Hathi Trust books, and one of their genres is poetry, which they extend to page level labels in later work (Underwood, 2014). Other recent work uses image classification approaches (Lorang et al., 2015), focuses on Australian newspapers (Kilner and Fitch, 2017) or is language-specific on a small collection (Tizhoosh et al., 2008).

These existing approaches cannot be cleanly applied to discover poems such as this poem about "Sweet Peas" that our algorithm identified in the middle of a gardening guide (Figure 1).



88 GARDEN GUIDE

considered a poor time to move them. When Iris clumps begin to choke themselves out by covering the ground so that young shoots have difficulty in establishing roots, they should be broken up and set in another place. Due to the prolificacy of German Iris this will be necessary every third year.

SWEET PEAS

The poet has a jingle upon Peas. He says:

"Peas along the border, Peas upon the lawn,
Peas against an eastern wall to welcome in the dawn.
Peas among the Roses, Peas behind the Pinks;
Peas to catch the western glow when evening sunlight sinks.
Peas upheld with Chestnut, Peas held up with Ash;
Peas asprawl on Hazel spray, Peas on Larchen brash.
Peas on stiff, unyielding wire, Peas tied up with string;
Peas upon the trellis work where Rambler Roses swing.
Oh! merry, merry, merry, are the gay Sweet Peas;
Plant them when and how you will, it's certain they will please."

It would appear that the answer to the question of where to plant Sweet Peas is, "Everywhere." But Sweet Peas should be planted on a well drained soil only, or one in which the excessive rains of Spring will not cause water to stand around the roots and start mildew. They do not like excessive shade, for the plants should make a sturdy growth. In the shade the growth is weak and spindly and but few flowers are produced.

Figure 1: A Poem printed in the middle of a Gardening Guide (Rockwell et al., 1917). This is the kind of "hidden" poetry our algorithm was designed to target.

Drawing inspiration and ideas from these works, we formulated the poetry identification problem: does a given scanned book page contain poetry on it?

Using a few thousand labeled pages as training data and only language-independent features, we developed a new model for poetry identification. This model is both effective (F1 = 0.83) and efficient (500,000 books/hour - single machine). It runs on DJVU-XML books from the Internet Archive.

Public Resources, Code, & Open Data

We released a variety of public resources. There is a dataset of our identification task as well as a JSON-formatted collection of 600,000 pages identified to contain poetry from a random selection of 50,000 books. Our model is available and our methodology can be found in more detail in my dissertation (Foley, 2019). Datasets: http://ciir.cs.umass.edu/downloads/ poetryCode & Model: https://github.com/jjfiv/poetry-identificationDemo: https://poetry.jjfoley.me

References

Ahmed, M. A. and Trausan-Matu, S. (2017). Using natural language processing for analyzing arabic poetry rhythm. In Networking in Education and Research (RoEduNet), 2017 16th RoEduNet Conference, pages 1-5. IEEE.

Alsharif, O., Alshamaa, D., and Ghneim, N. (2013). Emotion classification in arabic poetry using machine learning. International Journal of Computer Applications, 65(16).

Barros, L., Rodriguez, P., and Ortigosa, A. (2013). Automatic classification of literature pieces by emotion detection: A study on Quevedo's poetry. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, pages 141-146. IEEE.

Baumann, T., Hussein, H., and Meyer-Sickendiek, B. (2018). Style detection for free verse poetry from text and speech. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1929-1940.

Can, E. F., Can, F., Duygulu, P., and Kalpakli, M. (2011). Automatic categorization of ottoman literary texts by poet and time period. In Computer and Information Sciences II, pages 51-57. Springer.

Choi, K., Lee, J. H., Hu, X., and Downie, J. S. (2016). Music subject classification based on lyrics and user interpretations. Proceedings of the Association for Information Science and Technology, 53(1):1-10.

Foley, J. (2019). Poetry: Identification, Entity Recognition, and Retrieval. PhD thesis, University of Massachusetts.

Hamidi, S., Razzazi, F., and Ghaemmaghami, M. P. (2009). Automatic meter classification in persian poetries using support vector machines. In Signal Processing and Information Technology, 2009 IEEE International Symposium on, pages 563-567. IEEE.

Jamal, N., Mohd, M., and Noah, S. A. (2012). Poetry classification using support vector machines. Journal of Computer Science, 8(9):1441.

Kaur, J. and Saini, J. R. (2017). Punjabi poetry classification: The test of 10 machine learning algorithms. In Proceedings of the 9th International Conference on Machine Learning and Computing, pages 1-5. ACM.

Kesarwani, V. (2018). Automatic Poetry Classification Using Natural Language Processing. PhD thesis, Universit ?e d'Ottawa/University of Ottawa.

Kilner, K. and Fitch, K. (2017). Searching for My Lady's Bonnet: discovering poetry in the National Library of Australia's newspapers database. Digital Scholarship in the Humanities.

Kumar, V. and Minz, S. (2014). Multi-view ensemble learning for poem data classification using sentiwordnet. In Advanced Computing, Networking and Informatics-Volume 1, pages 57-66. Springer.

Lorang, E. M., Soh, L.-K., Datla, M. V., and Kulwicki, S. (2015). Developing an image-based classifier for detecting poetic content in historic newspaper collections. Technical report, University of Nebraska - Lincoln.

Lou, A., Inkpen, D., and Tanasescu, C. (2015). Multilabel subject-based classification of poetry. In The Twenty-Eighth International Flairs Conference.

Promrit, N. and Waijanya, S. (2017). Convolutional neural networks for thai poem classification. In

International Symposium on Neural Networks, pages 449-456. Springer.

Rakshit, G., Ghosh, A., Bhattacharyya, P., and Haffari, G. (2015). Automated Analysis of Bangla Poetry for Classification and Poet Identification. In Proceedings of the 12th International Conference on Natural Language Processing, pages 247-253.

Rockwell, F., Loveless, A., and Hottes, A. (1917). Garden Guide: The Amateur Gardener's Handbook. Internet Archive: https://archive.org/details/gardenguideamate00dickrich/page/88.

Tizhoosh, H. R., Sahba, F., and Dara, R. (2008). Poetic features for poem recognition: A comparative study. Journal of Pattern Recognition Research, 3(1):24-39.

Underwood, T. (2014). Understanding Genre in a Collection of a Million Volumes. Technical report, University of Illinois, Urbana-Champaign.

Underwood, T., Black, M. L., Auvil, L., and Capitanu, B. (2013). Mapping mutable genres in structurally complex volumes. In IEEE Big Data, 95-103.

Building from Bones: DH-WoGeM and Care in DH

Fabiola Delfin
Universidad Nacional Autónoma de México, Mexico

Quinn Dombrowski
qad@stanford.edu
Stanford University, United States of America

Isabel Galina Russell
igalina@unam.mx
Universidad Nacional Autónoma de México, Mexico

Tassie Gniady
ctgniady@iu.edu
Indiana University, United States of America

Rachel Hendery
r.hendery@westernsydney.edu.au
Western Sydney University, Australia

Henriette Roued-Cunliffe
University of Copenhagen, Denmark

Zhenya Samoilova
Film University Babelsberg "Konrad Wolf", Germany

Valerie Shepard
vshepard@saonet.ucla.edu
University of California, Los Angeles, United States of America

Deb Verhoeven
debver@ualberta.ca
University of Alberta, Canada

Claire Warwick
c.l.h.warwick@durham.ac.uk
Durham University, United Kingdom

Amanda Wilson Bergado
Stanford University, United States of America

Keywords: care, ethics, organizations, labor
Topics: Global, English, Contemporary, digital activism and advocacy, meta-criticism (reflections on digital humanities and humanities computing), Gender and sexuality studies, Humanities computing

Deb Verhoeven's provocative question at DH 2015 in Australia - "Has anyone seen a woman?" - continues to resonate within digital humanities. In the following years, ADHO conference organizers have been more attentive to ensuring the visible presence of women, and have put measures in place (e.g. arranging childcare) to make it easier for women to attend. However, conference participation by women and gender minorities1 remains limited (Eichmann et al. 2016), and attention to the needs and concerns of this demographic at the conference level has little impact on their day-to-day experiences working in digital humanities.

The DH-WoGeM (Women and Gender Minorities* in DH) group was formed in fall 2018 to provide a forum for women and gender minorities to talk openly about the challenges they face at work, and possible ways to work towards addressing these issues. At DH 2019, the DHWoGeM "embodied DH" poster (as presented by a skeleton in a dress) won the most innovative poster award, and following the conference, the group submitted an application to be recognized as an ADHO Special Interest Group (SIG). The group's goal is to take an intersectional, culturally appropriate approach to making explicit the challenges that women and gender minorities face when working in digital humanities, and advocating for concrete, actionable measures to better

support women and gender minorities in their local context, as well as at their participation in international collaboration.

The work of DH-WoGeM has taken many forms, including synchronous virtual "conversations", a workshop, and an edited volume. This poster will distill key takeaway points from the group's work to date, along with pointers to resources that attendees can use to reflect on ways to foster a more inclusive work environment. We will also use the poster to solicit responses to the first international survey that the group has undertaken, on work/life balance and caring responsibilities (e.g. childcare, caring for ill relatives or friends, self-care). This survey, which will be open during summer 2020, will collect anonymous information about how these factors have impacted the work and career trajectories of people engaged in doing DH work at all levels, from students through retirees, and in as many countries as possible. The survey is unprecedented in its attempt to gather an on-the-ground view of the experience of doing DH across the globe, and the ways that different cultural contexts mitigate and/or exacerbate the difficulty of maintaining work/life balance when engaged in DH work.

Bibliography

Eichmann, Nickoal, Jeana Jorgensen, Scott B. Weingart. "Representation at Digital Humanities Conferences (2000-2015)". http://scottbot.net/representation-at-digital-humanities-conferences2000-2015/

Verhoeven, Deb. "Has Anyone Seen a Woman?" July 2, 2015. http://debverhoeven.com/anyoneseen-a-woman/

* We use the term "gender minorities" here to encompass a range of identities other than cis-men and -women, including non-binary and trans*-identifying people.

Towards Paleographic Linked Open Data (PLOD): A general vocabulary to describe paleographic features

Timo Homburg
timo.homburg@gmx.de
Mainz University Of Applied Sciences

Keywords: Paleography, Linked Data, Vocabulary
Topics: Comparative (2 or more geographical areas), English, BCE-4th Century, 5th-14th Century, 15th-17th Century, data modeling, linked (open) data, Computer science, Philology

Introduction

Digital paleography has been emerging as a field of research since the beginning of the new century. Paleographers, describe how a text has been displayed, and collect information such as writing styles, contextual information and the epoch. Ciula (2017) gives a good summary about digital paleography and its practices. Challenges concerning digital paleography are discussed by Hassner et al. (2015) for writing systems and discussed more broadly in Stokes (2015). The main results of these publications is, that the community of digital paleographers,which was mainly focused on a few particular languages, sees to broaden its scope and to research more global models to represent paleographic features such as writing style, typographic features and anomalies in texts. This calls fora unified representation of paleographic features to which this publication would like to contribute by suggesting a core vocabulary for paleographic description of texts.

Related Work

Paleographic features can to a certain extent be represented in TEI/XML1 Wit-tern et al. (2009) (scribal shifts, scribal alterations) as highlighting annotations. For systematic non-alphabetic languages, encodings can be developed to represent the shape of the characters, e.g. for cuneiform and egyptian hieroglyphics Homburg (2019), van den Berg (1997)

In the linguistic community, linguistic linked open data (LLOD) McCrae et al.(2016) is very present and allows for tools such as BabelNet Navigli & Ponzetto (2012), a multilingual semantic network to translate words and texts based onsemantic content. This publication suggests to create such a semantic network for paleographic descriptions in order to formalize this part of research. Naturally, such a task cannot be done by a single scholar from a single field, so the publication begins by suggesting vocabulary contents documenting the structure of systematic scripts like cuneiform and to model relations between components of parts of scripts, the so-called core vocabulary. Especially structured scripts (cf. fig. 1) expose a variety of encoding schemes for e.g. Chinese Bishop & Cook (2003), Japanese Apel (2014). Those encodings have been created to modelfonts, but form an ideal basis to create the proposed unified vocabulary. Similar to the OliA ontologies Chiarcos & Sukhareva (2015) the author suggests to extend

this core vocabulary representing the structure of the script (figure 3) by language/script specific extensions. The concept will be presented using the example of cuneiform and egyptian hieroglyphics and builds upon the vocabulary shown in figure 2.

205 characters

Sign	Code point	Transliteration	Borger	PaleoCode	SVG
𒀀	U+12000	A	839	a-a:a	
𒀁	U+12001	A x A	845	a-:sa-:sa:sa-a:a	
𒀂	U+12002	A x BAD	840	a-b-w-a:a	
𒄰	U+122F0	TAB	209	b:b	
𒌋	U+12311	U2	490	B::B-a-a-a-a	

Figure 1. Example of the PaleoCodage encoding Homburg (2019): PaleoCodage represents a machine-readable way to describe highly structured scripts such as cuneiform.This structure, relations to other similar characters and scripts can be modelled using the proposed core vocabulary, while the language specific vocabularies would describe writing styles, shapes and stilistic characteristics of the respective script.

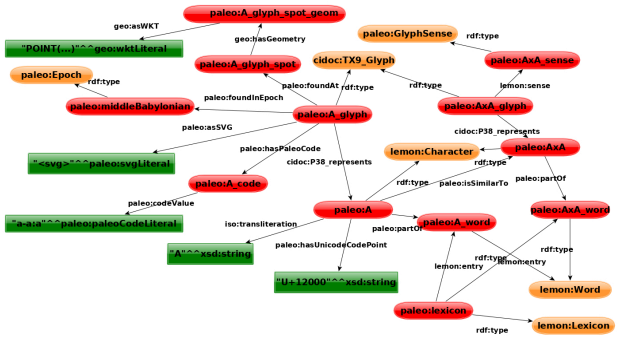


Figure 2:
Vocabulary describing two cuneiform glyphs connected to character representations, their finding spot, their assigned epoch, an assigned glyph sense (whichmay be distinct from the character/word sense) and possible serializations in SVG andas a PaleoCode. For other languages other encodings are possible.

```
paleo:A rdf:type lemon:Character .
paleo:A_code rdf:type paleo:Code .
paleo:A paleo:hasPaleoCode paleo:A_code .
paleo:A_code paleo:codeValue "a-a:a"^^paleo:paleoCodeLiteral .
paleo:A paleo:hasUnicodeCodePoint "U+12000"^^xsd:string .
paleo:A iso:transliteration "A"^^xsd:string .
paleo:A paleo:isSimilarTo paleo:AxA .
paleo:A paleo:isIncludedIn paleo:AxA .
paleo:A paleo:borgerNumber "839"^^xsd:string .
paleo:A_code paleo:asSVG "<div class='svgcontainer'>
<svg version='1.1' xmlns='http://www.w3.org/2000/svg'
xmlns:xlink='http://www.w3.org/1999/xlink' width='170'
height='80'><defs></defs><g><path fill='#000000' stroke='#000000'
paint-order='fill stroke markers'
d='M 5 20 L 15 20 L 10 30 L 5 20 M 10 30 L 10 60 M 15 20
 L 25 20 L 20 30 L 15 20 M 20 30 L 20 60 M 15 27 L 25 27
 L 20 37 L 15 27 M 20 37 L 20 67' stroke-miterlimit='10'
 stroke-dasharray=''></path></g></svg></div>"^^xsd:string .
paleo:A_code paleo:gottsteinCode "a3"^^xsd:string .
```

Figure 3: Excerpt: PaleoCodage Vocabulary describing a cuneiform sign. The sign's structure is described using a PaleoCode which could itself be described using Semantic relations. These relations allow to model the structure of script subelements with extensions for paleographic features

Bibliography

Apel, U. (2014), 'Kanjivg', kanji SVG dataset, Creative Commons BY-SA 3. 3Bishop, T. & Cook, R. (2003), A specification for cdl character descriptionlanguage, in 'Glyph and Typesetting Workshop'. 3Chiarcos, C. & Sukhareva, M. (2015), 'Olia-ontologies of linguistic annotation',Semantic Web 6(4), 379-386. 3Ciula, A. (2017), 'Digital palaeography: What is digital about it?', Digital Schol-arship in the Humanities 32(suppl_2), ii89-ii105. 1Hassner, T., Rehbein, M., Stokes, P. A. & Wolf, L. (2015), 'Computation andpalaeography: potentials and limits', Kodikologie und Paläographie im Dig-italen Zeitalter 3: Codicology and Palaeography in the Digital Age 3 p. 1.1Homburg, T. (2019), Paleo codage - a machine-readable way to describecuneiform characters paleographically, Utrecht, Netherlands.URL: https://dev.clariah.nl/files/dh2019/boa/0259.html 2, 1McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., De Melo,G., Gracia, J., Hellmann, S., Klimek, B., Moran, S. et al. (2016), The openlinguistics working group: developing the linguistic linked open data cloud,in 'Proceedings of the Tenth International Conference on Language Resourcesand Evaluation (LREC 2016)', pp. 2435-2441. 3Navigli, R. & Ponzetto, S. P. (2012), 'Babelnet: The automatic construction,evaluation and application of a wide-coverage multilingual semantic network',Artificial Intelligence 193, 217-250. 3Stokes, P. A. (2015), 'Digital approaches to paleography and book history: somechallenges, present and future', Frontiers in Digital Humanities 2, 5. 1van den Berg, H. (1997), 'Manuel de codage: A standard system for thecomputer-encoding of egyptian transliteration and hieroglyphic texts', Centrefor Computer-Aided Egyptological Research (CCER) . 2Wittern, C., Ciula, A.

& Tuohy, C. (2009), 'The making of tei p5', Literary andLinguistic Computing 24(3), 281-296. 2

This presentation presents first efforts to develop a vocabulary for paleographic features across different languages. It will feature a case study using cuneiform languages to infer relevant elements for a general vocabulary for paleography. The poster can be seen as a point of discussion and invitation for colleagues of similar and other fields to further develop the vocabulary which can be used to describe writing styles. When finished it is planned to be used among others to benefit machine learning tasks in computational lingustics as a paleographic desctiption of signs can act as a source of features which is rarely exploited in such scenarios.

The Digital MedievalistPostgraduate Subcommittee Networking, Research, and Open Collaboration

James Buffington Harr. III
jbharr@ncsu.edu
North Carolina State University, United States of America

Keywords: medieval, network, data visualization, paleography, history
Topics: Global, Europe, English, North America, 5th-14th Century, manuscripts description, representation, and analysis, scholarly editing and editions development, analysis, and methods, History, Literary studies

Digital Medievalist (digitalmedievalist.wordpress.com) is an international, interest-based virtual research community with a broad thematic focus that brings together academics and social scientists of varying status and promotes ways to jointly expand the scope of the individual disciplines under the umbrella of "medieval studies". During the poster session we would like to introduce the various initiatives of the Digital Medievalist Community and promote networking within the DH community. In particular, the presentation will outline the planned activities of the newly-founded Postgraduate Subcommittee - organization of joint panels, production of podcasts, general social media presence - and promote a peer-to-peer exchange.

The Petition Project: Visualizations and Representation of the 13th-Century Communication Network of Edward I

James Buffington Harr. III
jbharr@ncsu.edu
North Carolina State University, United States of America

Keywords: petitions, networks, communication, medieval, materiality
Topics: Global, English, 5th-14th Century, manuscripts description, representation, and analysis, network analysis and graphs theory and application, History, Law and legal studies

This paper examines the use of network theory to analyze data extracted from petitionary texts as a method to form prosopographical narratives of 13th century English society. The source material for this project is drawn from a sample of open-source digitized documents available from the British National Archives, Special Collections 8 (TNA, SC8), specifically the petitions from the late-13th-century addressed to Edward I.

"Empowered Minorities Differential Outcomes For Minorities Enjoying Kremlin Support"

Martha Brill Olcott
Michigan State University, United States of America

Frances Davis
Michigan State University, United States of America

Keelyn Obrien
Michigan State University, United States of America

Katya Trotskyi
Michigan State University, United States of America

Michael Downs
Michigan State University, United States of America

Bridget McBride
Michigan State University, United States of America

Keywords: USSR, language rights, Georgia, Baltic states, ethnic minorities
Topics: Comparative (2 or more geographical areas), English, 20th Century, digital libraries creation, management, and analysis, ethnographic analysis, History, Central/Eastern European Studies

This presentation is part of "Collapse and Rebirth: A Living Archive on the End of the USSR and Afterward (1985-1994)." We have already developed the prototype website for the project, using a combination of two platforms (Esri Arcgis Storymaps, and Omeka S.The site will support public commentary by eyewitnesses, suggestions for additional events to be featured, and offers of archival material for curation and inclusion. It will include bibliographic entries for each event, and will provide links to other archival materials and digital resources on the topic that are available on the internet. We are proposing a poster created by MSU faculty and student team members at DH2020 introducing second phase "stories" on the political movements and protests that enjoyed Kremlin support as a way to disadvantage other national minorities during the period 1989-1991 in several republics and autonomous regions of the USSR.

Women and Power in the Later Roman Empire: Diachronic and Geospatial Analysis Possibilities from TEI-Encoded Roman Legal Codes

JaShong King
jking113@uottawa.ca
University of Ottawa, Canada

Lydia Schriemer
lschr023@uottawa.ca
University of Ottawa, Canada

Keywords: roman, empire, law, tei, gis
Topics: Asia, Africa, Europe, English, BCE-4th Century, 5th-14th Century, database creation, management, and analysis, spatial & spatio-temporal analysis, modeling and visualization, History, Law and legal studies

This poster will highlight two different analytical possibilities from a robust TEI (Text-Encoding Initiative) XML (eXtensible Markup Language) encoding of an ancient source, the Theodosian Code, one of the foundational documents for Roman and, by extension, modern law. The first attempts to correlate laws affecting women to imperial policy; the second highlights the power projections of emperors through their travels. These two act as proofs-of-concept for a larger project, the analysis of Late Roman socio-political structures via a full digital editioning of the Theodosian Code.

Measuring and visualizing narrative structure in Jan Potocki's Manuscript Found in Saragossa

Scott Goldstein
scott.goldstein@mcgill.ca

McGill University

Keywords: data visualization
Topics: Global, English, 15th-17th Century, 18th Century, 19th Century, database creation, management, and analysis, text mining and analysis, Ethnography and folklore, Literary studies

Jan Potocki's Manuscript Found in Saragossa has been described as a classic of world literature and is arguably the most complex novel ever written in terms of its nested and interconnected structure of narratives. Because of its sheer complexity, devices have been introduced to help the reader make sense of the overall structure, but none are interactive, and some are based on the more well-known 1965 film of the same name, which differs from the novel. The current project is the generation of an open data set based on the only public domain version of the novel, Edmund Chojecki's Polish translation. The data set is a database of every narrative segment of the novel, including the name of the story, its location within the nested structure of the novel, and the full text, which will allow for a variety of analyses including word counts, topic modeling, and other narratological techniques.

Integrating Historical Maps and Documents through Geocoding - Historical Big Data for the Japanese City of Edo

Asanobu Kitamoto
kitamoto@nii.ac.jp
ROIS-DS Center for Open Data in the Humanities, Japan National Institute of Informatics

Shoko Terao
AMANE LLC.

Misato Horii
AMANE LLC.

Hiroshi Horii
AMANE LLC.

Chikahiko Suzuki
ROIS-DS Center for Open Data in the Humanities, Japan National Institute of Informatics

Keywords: data integration, geocoding, historical maps, Japanese city of Edo, urban space
Topics: Asia, English, 19th Century, digital libraries creation, management, and analysis, spatial & spatio-temporal analysis, modeling and visualization, Asian studies, Geography and geo-humanities

1. Historical Big Data for the Urban Space

City of Edo has been the capital of Japan and is known to be one of the largest city in the world since 17th Century. To answer research questions on historical urban space, such as human activities and environmental effects, historical documents should be integrated by place, time, person and other entities to turn small facts into a collection of structured data for historical big data analysis [1]. Related work includes Pelagios [2], which studies historical gazetteers and georeferencing of old maps to reconstruct the geographic space, and European Time Machine [3], which aims at integrating historical entities to reconstruct the urban space in European cities. Our approach could also be called as Edo Time Machine.

2. Integration of Historical Sources through Geocoding

Toponyms are described in many variations, especially on historical documents before the standardization of the address system. Hence a location-based historical database requires the shared address system, or the standard gazetteer, for toponym-based integration. Major challenges in toponym-based integration is variation and disambiguation of toponyms, and a question in this paper is how machine-based geocoding can deal with these challenges.

3. Dataset

Edo Map Dataset: The dataset covers place names extracted from "Edo Kiriezu" (Owariya version) [4], a pre-modern map of Edo published from 1849 through 1863 in the form of 32 sheets. It contains not only addresses but also POIs (Point of Interests) such as bridges and temples [5].Edo Shopping Dataset: The dataset covers shops and restaurants extracted from "Edo Kaimono Hitori Annai" [6], a pre-modern shopping guide published in 1824 about 2600 shops and restaurants in Edo with the shop name, category, address and logo [7].

To create the dataset, we took advantage of IIIF (International Image Interoperability Framework), which allows interoperable image delivery in the humanities, and IIIF Curation Platform (ICP) [8,9,10], which is an

open source software suite developed by our group to create the collection of a part of images across organizations. As a result, we created the dataset of 6418 place names from 22 sheets out of 32 sheets (Figure 1), and the dataset of 2454 shops from the whole book (Figure 2).

Figure 1: Edo Kiriezu, the sheet of Yotsuya area. Red markers show extracted place names (Total 335).

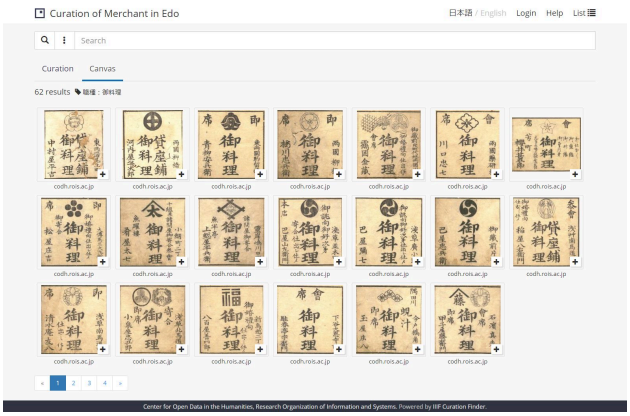


Figure 2: Edo Kaimono Hitori Annai. A search result for restaurants (Total 62).

4. Experimental Results

Table 1 shows the result of matching between an entry in the gazetteer and a shop address (1034 unique addresses). In addition to exact match, we tested three other approaches; matching from the first character (forward match), matching from the last character (backward match), and matching a part of the address string (partial match). Table 1 shows that exact match was successful for about 21% (212/1034). Among the 212 successful cases, 49 addresses need disambiguation within a sheet and 15 needs disambiguation across sheets. Disambiguation within a sheet, however, is usually not a critical issue because, under the block-based, instead of street-based, Japanese address system, it usually means multiple neighboring blocks. Future work includes georeferencing coordinates between old maps and the present map, and analyzing relationship between the geographic distribution of businesses and human activities in the urban space.

	Exact match	Forward match	Backward match	Partial match
Success cases	148	294	320	71
Disambiguation within the sheet	49	147	212	53
Disambiguation across sheets	15	129	127	38

Table 1: Matching 1034 unique addresses in the shopping guide against place names in the gazetteer. Note that some categories are not mutually exclusive.

References

[1] Kitamoto, A., Ichino, M., Suzuki, C., Clanuwat, T (2018). Historical Big Data: Reconstructing the Past through the Integrated Analysis of Historical Data. Eighth Conference of Japanese Association for Digital Humanities (JADH2018). pp. 67-69.

[2] Isaksen, L., Simon, R., Barker, E. T., & de Soto Cañamares, P. (2014). Pelagios and the emerging graph of ancient world data. In Proceedings of the 2014 ACM conference on Web science, pp. 197-201.

[3] Time Machine Europe, https://www.timemachine.eu/, Accessed on June 15, 2020.

[4] Owariya (1849-1862), Edo Kiriezu, National Diet Library, https://www.ndl.go.jp/landmarks/edo/, Accessed on June 15, 2020.

[5] Center for Open Data in the Humanities, Edo Maps Beta, http://codh.rois.ac.jp/edo-maps/, Accessed on June 15, 2020.

[6] Nakagawa Gorozaemon (1824). Edo Kaimono Hitori Annai, Dataset of Pre-modern Japanese Text (photographed by National Institute of Japanese Literature, archived by Ajinomoto Foundation For Dietary Culture), doi: 10.20730/100249503.

[7] Suzuki, C., Curation of Merchant in Edo, http://www.ch-suzuki.com/edoshop/finder/?lang=en, Accessed on June 15, 2020.

[8] Kitamoto, A., Homma, J., Saier, T. (2018) IIIF Curation Platform: Next Generation IIIF Open Platform Supporting User-Driven Image Sharing. Proceedings of IPSJ SIG Computers and the Humanities Symposium 2018. pp. 327-334 (in Japanese).

[9] Kitamoto, A. (2019) IIIF Curation Platform: Creating and Sharing Virtual Image Collection on a Global Scale. 2019 International Conference: Glocal Humanities in the Era of Hyperconnectivity. 6 pages.

[10] Center for Open Data in the Humanities, IIIF Curation Platform, http://codh.rois.ac.jp/icp/, Accessed on June 15, 2020.

THE DIGITAL TURN IN AFRICA - ACADEMIC LIBRARIES AS INTERSECTIONS FOR DIGITAL HUMANITIES INITIATIVES

Joseph Ssebulime
josephssebulime1@gmail.com
Makerere University, Uganda

Johann Van Wyk
University of Pretoria, South Africa

Keywords: DIGITAL TURN IN AFRICA, ACADEMIC LIBRARIES, INTERSECTIONS, DIGITAL HUMANITIES INITIATIVES
Topics: Afric, aEnglish, 19th Century, 20th Century, Contemporary, data publishing projects, systems, and methods, scholarly editing and editions development, analysis, and methods, Library & information science, Literary studies

At Makerere University, Uganda, the Library is responsible for meeting the needs of scholars. The role of human scientists is to investigate and answer existing and emerging research questions in the humanities.

The existing data sources for the human scientists at Makerere University are increasing rapidly. The wealth of data flowing from these are a mixture of structured and unstructured data, that come in a variety of formats, some of which can contain noise and can be partially incomplete.

Dealing with this growth and complexity of data, necessitates the use of various digital techniques by human scientists, however, would require guidance and facilities - an area where academic libraries can play a crucial role.

Academic libraries have become carrefours / intersections for digital humanities as partners in the research process. This poster will present how Makerere Library is at the center of supporting Digital Humanities Initiatives through the provision of various services.

Zemsky Educational Establishments: Development of the Database

Iliana Damirovna Ismakaeva
ilianaismakaeva9@gmail.com
National Research University Higher School of Economics, Russian Federation

Keywords: zemstvo, local self-government bodies, literacy, primary education, Russian empire
Topics: Europe, English, 19th Century, 20th Century, database creation, management, and analysis, Education/ pedagogy, History

The process of transformation, which began in the Russian Empire in the second half of the XIX century, as one of the important social problems put forward the question of literacy. The problem of literacy in Russia intensified and became more prominent after the abolition of serfdom in 1861 when peasants were granted personal freedom and some civil rights. The elimination of illiteracy was one of the most important conditions for the further development of the country. The spread of literacy in the Russian Empire in the late XIX-early XX centuries was largely associated with the activities of local self-government bodies. For them one of the most important areas of work was public education. The well-known fact is that the local self-government reform was carried in 1864. In Russian, these local self-government bodies were called Zemstvo.

These local self-government bodies did not have political functions, they dealt with building roads, improving streets, opening and maintaining schools and churches.

My field of research concerns the problems of the local self-government bodies activity in the Perm province (Russian Empire). Perm Zemstvo was established in 1870. At that time, Perm Zemstvo founded more than 2000 educational establishments which contributed to the development of literacy.

The Perm local self-government separately considered the issues of the creation and development of the indigenous non-Russian schools. The largest non-Russian population in the Perm province was Permian, Tatar and Bashkir (according to the population census of 1897 the Permian population was 92642, Tatar-46711, Bashkir-85395). The Perm local self-government founded specific non-Russian schools for these peoples.

I am collecting information about each of the educational establishments of Zemstvo in Perm. Currently I am working on database creation. This database allows keeping and processing information about the educational establishments such as names of these educational establishments, location, the time of their work, the number and name of teachers, subjects and textbooks.

This database will be posted publicly on the Internet. Teachers can use this database in the educational process for schoolchildren and students, as well as organize educational projects together with museums, archives, libraries and educational institutions. Researchers will be able to use the database to conduct a local history research and research in the field of public history. Students, schoolchildren and researchers are likely find in this database information about their schools which in the late 19th and early 20th centuries functioned as schools of Zemstvo.

Supervised Learning and the Indeterminacy of Disordered Genres

Julian Schroeter
julian.schroeter@uni-wuerzburg.de
University of Wuerzburg, Germany

Keywords: supervised learning, genre theory, genre history, novella, aesthetics
Topics: Europe, English, 19th Century, cultural analytics, text mining and analysis, Cultural studies, Literary studies

The poster puts forward the idea of interpreting accuracy scores of classification tasks as a measure of the psychological degree of looseness of historical genre concepts. Based on a more comprehensive corpus, this poster shall introduce an unconventional way of applying supervised machine learning to describe the peculiar order of disordered genres in three short steps: Firstly, the notion of "disordered genres" has to be explicated. Secondly, results of classification tasks are presented, and, thirdly, the idea of interpreting strong as well as weak validation scores as a measure of the degree of the psychological manageability of genre concepts readers generate from reading is put forward. If historical reader responses correspond to the statistical results, validation scores of supervised learning tasks can be interpreted as a metric that measures the degree of looseness of historical genre concepts in general.

Taming the Data: Web-Scraping and De-Duplicating Messy Multilingual Philosophy Corpora

Raluca A. Tanasescu
r.a.tanasescu@rug.nl
University of Groningen, Netherlands, The

Cristian A. Marocico
c.a.marocico@rug.nl
University of Groningen, Netherlands, The

Keywords: web crawling, text analysis, data cleaning, deduplication, corpus expansion
Topics: Europe, English, 15th-17th Century, 18th Century, information retrieval and querying algorithms and methods, text mining and analysis, Humanities computing, Philosophy

This poster presents a technical report and a method for corpus expansion in the humanities, with an application to early modern philosophy, alongside a case study of dealing with heavy data redundancy in several Latin, English, and French title corpora. It enlarges on the steps taken during the initial stages of a data-intensive research project that aims to go beyond established writers and views in natural philosophy between 1600 and 1800 and it reflects on the collaboration between a humanist and a data scientist with respect to web-scraping and redundant multilingual data taming in Python.

Cartesianism as a Social Epidemic: A Network Analysis Approach

Paolo Rossini
prossini1991@gmail.com
University of Leiden, Netherlands, The

Keywords: Descartes, Network analysis, Social epidemics
Topics: Europe, English, 15th-17th Century, network analysis and graphs theory and application, Philosophy

The aim of this paper is to give a concrete example of how network analysis can be applied to study the dissemination of new ideas throughout history. The father of modern philosophy, René Descartes (1596--1650), is taken as a case study. To trace the rapid spread of Descartes' ideas across early modern Europe, I will rely on ePistolarium, a web archive of more than 20,000 digitized letters. Using the metadata and transcriptions of ePistolarium, I will model a network in which the edges will represent the letters mentioning Descartes and the nodes will represent the people connected by those letters. Studying the topology of the network, I will be able to determine which actors were more involved in spreading Descartes' ideas and the routes along which information about Descartes travelled.

Victorian Poverty in Print and Digital Forms: Digitally Representing the Multimodal Publication History of Henry Mayhew's London Labour and the London Poor

Jenna M Herdman
jenna.herdman@carleton.ca
Carleton University, Canada

Keywords: scholarly editing, hypertext, Victorian studies, print culture
Topics: Europe, English, 19th Century, electronic literature production and analysis, scholarly editing and editions development, analysis, and methods, Book and print history, Literary studies

This project digitally remediates selections from Henry Mayhew's seminal survey of poverty, London Labour and the London Poor (weekly serial, 1850-52; 2 vols, 1851; 4 vols, reissued with additions, 1861). Print culture offers rich opportunities for digitization, and I engage with current conversations in Victorian studies about developing sustainable, scholar-driven digital resources produced outside of commercial ventures like GALE and ProQuest (Stauffer; Fyfe). For example, Felluga argues for resisting skeuomorphic forms of knowledge production by capitalizing on new technologies (44), while Wisnicki points to the success of nineteenth-century digital archives which emphasize materiality, are user-friendly, and are sustainable with little funding (984). I engage with work on the potential of digitization as a means of navigating and preserving Victorian print media (Mussell), on canonical scholarship of hypermedia and digital textualities (Landow; Bolter; Gaggi; McGann; Hayles), and on recent work on the hypermedia reading and cognitive hypertext comprehension (Ryan; Shang; Medina-Medina et al.) to examine how interactive media offers a model of digitally representing Victorian print culture which resists the form of the printed book. The theoretical framing of this project addresses the following questions: 1) What potential is there for interactive hypertext to represent the materiality, dissemination, and readerships of serialized print texts? 2) How do Victorian texts speak to the concerns of digital publishing and materiality in a public digital humanities context?

As a seminal text in the history of urban poverty, London Labour has been subject to continuous remediation and repurposing. It was attributed to Henry Mayhew, but was the collaborative product of multiple actors, including multiple journalists, editors, publishers, readers, and the London poor themselves. The multimodal remediation of pieces of this collaborative work as a newspaper column, independent serial, expensive volume set, stage play, reading series, 'unfashionable' novel, and twenty-first century selected print editions, destabilizes readings of London Labour as an authoritative encyclopedic resource of nineteenth-century culture and poverty. Its volume and irregular production history make it difficult to engage with (Roddy et al. 482; Schroeder), a

common problem when studying periodicals (Robson). In its current digitized forms on Google Books and Internet Archive, London Labour is represented using facsimile scans of its four volumes, which, despite being comprehensive and searchable, are not subject to bibliographic control, and feature unreliable OCR. However, London Labour invites a nonlinear and interconnected model of reading that speaks to, and even anticipates, hypertextual forms of reading. My project features twenty-five of Mayhew's articles, selected according to two pressing themes: first, waste collection and recycling, and second, the publication history and the materiality of books in London Labour. These selections contain scholarly annotations and are hyperlinked in connecting pathways. They can be read chronologically, or by selecting keywords and thematic clusters. The selections also compare plain-text transcripts to nineteenth-century facsimiles, emphasizing the text's uneven production. Drawing from Scanlan's reading of the poor as producers circulating in an alternative waste economy, and Price's work on the material cycles of paper and books in Mayhew's work, my project interrogates the following carrefours: first, the intersections between Mayhew and the street-folk in the urban space of Victorian London represented in London Labour; and, more broadly, the intersection between Victorian print and the digital humanities. Receiving feedback will position me to expand this open-access resource as a collaborative project, and to develop its pedagogical potential for training students in digital methods and tools through the lens of print studies.

Works Cited

Bolter, J. David. Writing Space: Computers, Hypertext, and the Remediation of Print, 2nd ed., Lawrence Erlbaum, 2001.

Felluga, Dino Franco. "BRANCHing out: Victorian studies and the digital humanities." Critical Quarterly, vol. 55, no. 1, 2013, pp. 43-56.

Fyfe, Paul. "Access, Computational Analysis, and Fair Use in the Digitized Nineteenth-Century Press." Victorian Periodicals Review, vol. 51, no. 4, 2018, pp. 716-737.

Hayles, N. Katherine. (2002). Writing Machines. MIT Press, 2002.

Landow, George P. Hypertext 3.0: Critical Theory and New Media in an Era of Globalization. Johns Hopkins U P, 2006.

Mayhew, Henry. London Labour and the London Poor. 4 vols. London: Griffin, Bohn, and Company, 1861. Google Books, https://books.google.ca/books?id=pcoTAAAAQAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false. Accessed 10 Oct. 2019.

Mayhew, Henry. London Labour and the London Poor. 4 vols. London: Griffin, Bohn, and Company, 1861. Internet Archive, https://archive.org/details/in.ernet.dli.2015.37797/page/n9. Accessed 10 Oct. 2019.

McGann, Jerome. Radiant Textuality: Literature after the World Wide Web. Palgrave, 2001.

Medina-Medina, Nuria; Molina-Ortiz, Fernando; Padilla-Zea, Natalia; Cabrera-Cuevas, Marcelino; Garcia-Cabrera, Lina; Parets-Lloca, Jose. "Evolutionary authoring tool for adaptive hypermedia with multimodal navigation." New Review of Hypermedia and Multimedia, vol. 19, 2013, pp. 84-111.

Mussell, James. The Nineteenth-Century Press in the Digital Age. Palgrave Macmillan, 2012.

Price, Leah. "The Book as Waste: Henry Mayhew and the Fall of Paper Recycling." How to Do Things with Books in Victorian Britain, Princeton U P, 2012, pp. 219-257.

Robson, Catherine. "How We Search Now: New and Old Ways of Digging up Wolfe's 'Sir John Moore.'" Virtual Victorians: Networks, Connections, Technologies. Eds. Veronica Alfano and Andrew Stauffer. Palgrave Macmillan, 2015, pp. 11-28.

Roddy, Sarah; Strange, Julie-Marie; Taithe, Bertrand. "Henry Mayhew at 200 - the 'Other' Victorian Bicentenary." Journal of Victorian Culture, vol. 19, no. 4, pp. 481-496.

Ryan, Marie-Laure. Avatars of Story: Narrative Modes in Old and New Media, U of Minnesota P, 2006.

Scanlan, John. "In Deadly Time: The lasting on of waste in Mayhew's London." Time & Society, vol. 16, no. 2/3, 2007, pp. 189-206.

Schroeder, Janice. "The Publication History of Henry Mayhew's London Labour and the London Poor." BRANCH: Britain, Representation and Nineteenth-Century History. Ed. Dino Franco Felluga. Extension of Romanticism and Victorianism on the Net, http://www.branchcollective.org/?ps_articles=janice-schroeder-the-publishing-history-of-henry-mayhews-london-labour-and-the-london-poor. Accessed 8 Oct. 2019.

Shang, Hui-Fang. "Online metacognitive strategies, hypermedia annotations, and motivation on hypertext comprehension." Educational Technology & Society, vol. 19, no. 3, 2016.

Stauffer, Andrew. "Introduction." Virtual Victorians: Networks, Connections, Technologies. Eds. Veronica Alfano and Andrew Stauffer. Palgrave Macmillan, 2015, pp. 1-8.

Wisnicki, Adrian S. "Digital Victorian Studies Today." Victorian Literature and Culture, vol. 44, 2016, pp.

Non–Linear Temporality in The Sound and the Fury -- A Digital Humanities View

Ruifang Li
nora_li@163.com
Shanghai International Studies University

Keywords: The Sound and the Fury, non-Linear chronology, constraint logic programming, space of contested meanings, simplex
Topics: GlobalEuropeEnglishNorth America20th Century, Contemporary, spatial & spatio-temporal analysis, modeling and visualization, text encoding and markup language creation, deployment, and analysis, Humanities computing, Literary studies

This paper uses constraint logic programming to analyze the timeline of The Sound and the Fury. The basic idea is to encode inter-temporal references into inequalities, and input these into a computer program. The computer then tests possible years when each event could have occurred, while satisfying each of the constraints.

If the constraints are inconsistent, then the program will identify the 'minimal conflict set'. If the constraints can be satisfied, then the program will output all possible orderings of events that are consistent with the constraints.

Instead of using computers to identify a unique 'correct' interpretation, this paper attempts to model the space of contested meanings within the text.

In linear programming, such a space is called a simplex. With large numbers of constraints, the simplex is an n-dimensional surface far beyond human visualization -- this monstrous geometrical object is the space of possible meaning for Faulkner's chronology.

The semantics of the U.S. street map

Benjamin M Schmidt
bs145@nyu.edu
New York University, United States of America

Keywords: word2vec, visualization, maps as texts as data
Topics: English, North America, 18th Century, 19th Century, 20th Century, spatial & spatio-temporal analysis, modeling and visualization, text mining and analysis, Geography and geo-humanities, History

This poster reappropriates tools of textual analysis into spatial humanities research by treating the road network of the United States as a single, massive text (of street names) read by navigators every day, and applying the increasingly common steps of representation learning to them; a high dimensional embedding using word2vec that captures forms of inter-relation between names, and a low dimensional re-embedding using UMAP for visualization.

The resulting visualization opens a useful space between the spatial and textual humanities that looks to see how the evolving patterns of distant reading may be useful in thinking about the landscape as a text in the most literal way possible.

Preservation and Ethical Re-Use of Crowdsourced Citizen Scholarship

Allyssa Guzman
allyssa.guzman@austin.utexas.edu
University of Texas at Austin, United States of America

Albert A. Palacios
aapalacios@austin.utexas.edu
University of Texas at Austin, United States of America

Keywords: metadata, crowdsourcing, transcription, digital archives, preservation
Topics: South America, English, North America, Contemporary, crowdsourcing, metadata standards, systems, and methods, Library & information science

Cultural repositories are increasingly tapping into the collective expertise of an interested public via crowdsourcing to improve the intellectual accessibility of handwritten sources and nourish their archival records . However, few are discussing the subsequent preservation and ethical re-use of these citizen contributions in the archival record and scholarship. We do not consider citizen contributors as simply a "sourced crowd;" they are often experts in the subject matter represented in these primary sources, and their collective scholarship (i.e. transcriptions, translations,

indexes) is worthy of preservation. Our goal is to configure workflows for the preservation of these collectively-produced resources. We believe that every individual who contributes to the interpretation and understanding of our collections should be given credit for their efforts in the archival record and in scholarship. Based on ethical and practical considerations, our aim is to develop use statements and citation guides for scholars re-using citizen contributions.

Digital Humanities among the iSchools: An Analysis of DH Courses

Chris Alen Sula
csula@pratt.edu
Pratt Institute, United States of America

Xiaoguang Wang
Wuhan University, China

Heejin Park
Hansung University, Seoul, South Korea

Claudia Berger
Pratt Institute, United States of America

Keywords: curriculum, pedagogy, information studies
Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, curricular and pedagogical development and analysis, Informatics, Library & information science

This poster presents an analysis of current DH courses at the iSchools, an international group of 106 schools, colleges, and departments that are responsible for training many information professionals working across libraries, archives, museums, and cultural heritage organizations-institutions that have been discussed as key sites of DH work and partners for collaboration. This study draws on Spiro's (2011) methodology, which examines assignments, readings, media types, key concepts, and technologies, in an attempt to characterize the "hidden curriculum" found throughout DH courses. In addition, we attend to issues of global similarities and local differences among the iSchool courses, as well as DH more broadly. In presenting this poster at the conference, we hope to raise awareness of unique aspects of DH education within iSchools and invite discussions with others about education and training in the field, potential partnerships, and pedagogical approaches.

Evaluating a Machine Learning Approach to Identifying Expressive Content at Page Level in HathiTrust

Nikolaus Nova Parulian
nnp2@illinois.edu
HathiTrust Research Center, School of Information Sciences, University of Illinois at Urbana-Champaign, USA

Kristina Hall
HathiTrust, University of Michigan, USA

Ryan Dubnicek
rdubnic2@illinois.edu
HathiTrust Research Center, School of Information Sciences, University of Illinois at Urbana-Champaign, USA

Yuerong Hu
yuerong2@illinois.edu
HathiTrust Research Center, School of Information Sciences, University of Illinois at Urbana-Champaign, USA

J. Stephen Downie
jdownie@illinois.edu
HathiTrust Research Center, School of Information Sciences, University of Illinois at Urbana-Champaign, USA

Keywords: machine learning, HathiTrust, copyright
Topics: EnglishNorth America20th Century, Contemporary, artificial intelligence and machine learning, copyright, licensing, and permissions standards, systems, and processes, Computer science, Library & information science

Evaluating a Machine Learning Approach to Identifying Expressive Content at Page Level in HathiTrust

Nikolaus Parulian1, Kristina Hall2, Ryan Dubnicek1, Yuerong Hu1, Stephen Downie1

1 HathiTrust Research Center, School of Information Sciences, University of Illinois at Urbana-Champaign

2 HathiTrust, University of Michigan

Introduction

HathiTrust [1] fully provides scanned images, plain text and metadata in support of their mission to contribute to research, scholarship and the sharing of human knowledge. Since facts, unlike expressive content, are exempt from copyright [2], this project seeks to use machine learning approaches to evaluate how often expressive content appears in the first 20 pages of a given HathiTrust volume, with an eye to potentially making this data open.[3]

Information contained in the first 20 pages of a volume can be useful to scholars. For example, the title page, table of contents, or acknowledgment page may contain useful information to understanding the volume. However, it is likely that some volumes include materials that have copyright protection in this same range. Some observed copyrighted materials in this page range are illustrations or even the main text itself [4]. One method to understand if expressive content is exposed in the first 20 pages would require manual page labeling, which is time-intensive. A machine learning approach is more efficient and could be well-suited to this type of prediction task [5], and we seek to answer these research questions:

- Can we develop a machine learning approach to help detect expressive contents in the first 20 pag es of HathiTrust volumes?
- How reliably does this approach match manual labeling data?

Methodology

Providing a high-quality dataset for training the machine learning model is essential, and human expertise is required. We manually sampled 900 volumes from HathiTrust and labeled each of the first 20 pages: either as 'factual' for a page with contents lacking creative expression and 'creative' if there is protected material on the page. Then we developed a workflow to use the statistical features of the page from the HathiTrust Research Center (HTRC) Extracted Features Dataset [6] as additional data to train our model. The features used included: token and line counts, tokens per line, and begin and end line characters.

Using the features above, we trained and compared four basic classification models on our feature set: Random Forest, Logistic Regression, Support Vector Machine, and Stochastic Gradient Descent [7]. Through this comparison we hope to both find the most accurate model as well as generally evaluate if a machine learning approach can be accurate for this task. The preliminary results of our prediction model can be seen in Figure 1.

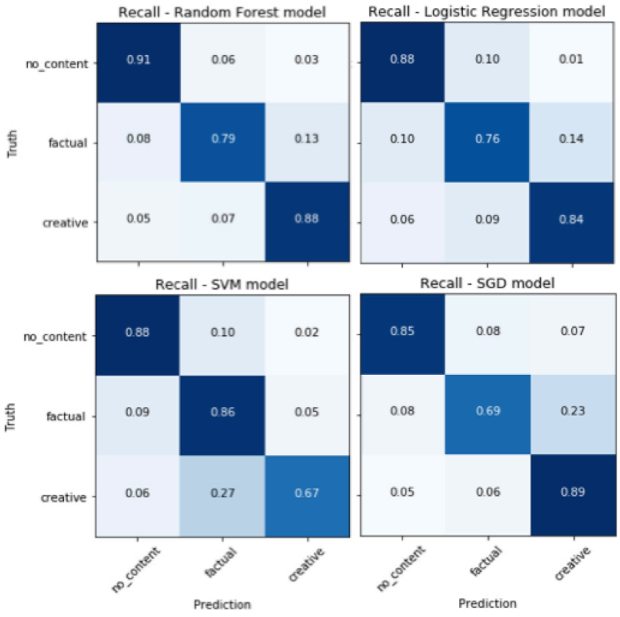


Figure 1: Confusion Matrix for four models for predicting creative (protected) content

Conclusion and Future Work

Results suggest that the Random Forest model performs best in both accuracy for predicting all labels (86%) and recall (0.88) for predicting creative content. For this project, we give more attention to the recall on the 'creative' label because a false negative on this label is a less desirable outcome. Future goals of this work are: to pilot different methods that can increase confidence in determining creative content, such as deep learning and utilizing page text, and to increase the scope of this prediction beyond our test set to a larger set of HathiTrust volumes.

References

1. https://www.hathitrust.org/
2. Justia Law. "Feist Pubs., Inc. v. Rural Tel. Svc. Co., Inc., 499 U.S. 340 (1991)." Accessed October 14, 2019.
3. https://supreme.justia.com/cases/federal/us/499/340/."Operationalizing 'Non-Consumptive' Fair Use to Revolut... | HathiTrust Digital Library." Accessed October 14, 2019. https://www.hathitrust.org/blogs/perspectives-from-hathitrust/operationalizing-non-consumptive-fair-use-to-revolutionize.

4. Example volume containing a mix of expressive and factual contents within front matter: https://catalog.hathitrust.org/ Record/005323091.
5. Lara McConnaughey, Jennifer Dai and David Bamman (2017), "The Labeled Segmentation of Printed Books" (EMNLP 2017)
6. Boris Capitanu, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, J. Stephen Downie (2016).The HathiTrust Research Center Extracted Feature Dataset (1.0) [Dataset]. HathiTrust Research Center, http://dx.doi.org/10.13012/J8X63JT3.
7. Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

The Digital Analysis of Syriac Handwriting (DASH) Project: Augmenting Manuscript Studies via Interactive Scriptcharts and IIIF

Michael Penn
Stanford University, United States of America

Vijoy Abraham
Stanford University, United States of America

Scott Bailey
scottbailey@stanford.edu
Stanford University, United States of America

Peter Broadwell
broadwell@stanford.edu
Stanford University, United States of America

Simon Wiles
simonwiles@stanford.edu
Stanford University, United States of America

Keywords: paleography, Syriac Christianity, digital manuscripts, IIIF
Topics: Asia, Africa, Europe, English, 5th-14th Century, image processing and analysis, manuscripts description, representation, and analysis, History, Theology and religious studies

The DASH digital paleography project uses ancient manuscripts written in the Aramaic dialect of Syriac as a case study for exploring how recent advances in the digital analysis of handwriting can help scholars better ascertain a manuscript's provenance, identify manuscripts written by the same scribe, and trace out the chronological development of ancient scripts.

At the core of the project is a curated manuscript database of metadata records describing 154 early, securely dated Syriac manuscripts, 1,432 digitized page images from these manuscripts, and the on-page coordinates of 67,242 individually identified letter images. An administrative interface with accompanying curation scripts enables project participants to manage data about the manuscripts and letter examples, and allows investigators to identify new letter examples.

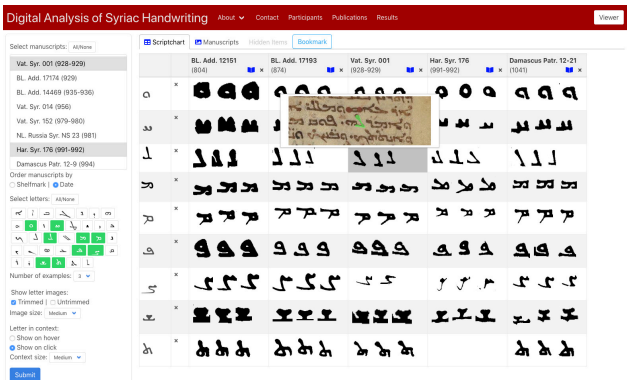


Fig. 1: The DASH interactive scriptchart app, displaying the letter-in-context feature

The interactive online manuscript viewer and customizable scripchart generator developed for this project (accessible at dash.stanford.edu) enables scholars to select any or all of the available manuscripts and Syriac letter forms for inclusion in a dynamically generated scriptchart. The scriptchart interface provides options to display up to the top five most highly rated instances of each selected letter on each manuscript at multiple zoom levels, and to view the letters in an excerpted, binarized form or highlighted within their original context on the manuscript page, with a variably sized context window around each letter instance to allow examination of its connections to neighboring letters. The chart facilitates dynamic reordering and hiding of manuscript columns and letter rows, providing scholars with an efficient means to propose and evaluate alternative localizations and chronologies for letter forms and manuscripts. In

addition, the interface includes a URL-based bookmark function so that a researcher can easily share a particular scriptchart configuration with others.

The project source code is freely available on Github, and data regarding the coordinates of the identified letters on specific manuscript pages are available via an API endpoint. By agreement with the manuscripts' host institutions, the interface also makes all manuscripts available for multi-up, zoomable viewing and navigation via embedded Mirador viewers that receive high-resolution images from the hosts' servers using the International Image Interoperability Framework (IIIF).

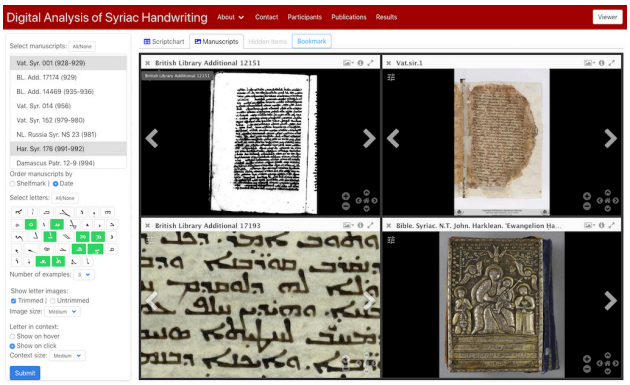


Fig. 2: Browsing multiple manuscripts in the DASH IIIF viewer

The database and interactive scriptcharts have facilitated studies of manuscripts that employ multiple classes of letters, with results that strongly suggest that the most typical way of classifying Syriac script as either exclusively Estrangela or Serto does not work particularly well for early manuscripts. Researchers also found that Estrangela letter forms were used as a "prestige" script and the frequency of their use showed a strong correlation with the genre of a given manuscript.

Although the initial goal of the DASH project is to advance the understanding of Syriac Christianity, its larger contribution is providing a platform and a model for similar ventures in other languages such as Arabic, Greek, Hebrew, Latin, and Sanskrit -- thereby positioning it at the crossroads of digital paleography.

References

"Challenging the Estrangela / Serto Divide: Why the Standard Model of Syriac Scripts Just Doesn't Work," Hugoye: Journal of Syriac Studies 21.1 (2018): 43-80. Kristina Bush, Michael Penn, R. Jordan Crouser, Nicholas Howe, and Shuangxia Wu.

"The Early History of the Syriac Script: A Reassessment." Journal of Semitic Studies 45.1 (2000): 55-67. J. F. Healey.

"Speculative Practices: Utilizing Infovis to Explore Untapped Literary Collections." IEEE Transactions on Visualization And Computer Graphics 22.1 (2016): 429-438. U. Hinrichs, S. Forlini, and B. Moynihan.

"The 'Emergency' Olaph: How Big Data About Little Things Changes Our View of Ancient Scribes" (in process). Michael Penn, Shuangxia Wu, Kristina Bush, R. Jordan Crouser, and Nicholas Howe.

Russian Across Borders Towards Virtual Language-Centric DH Working Groups

Sarah McEleney
sarmceleney@gmail.com
The University of Virginia, United States of America

Kaylin Land
kaylin.land@mail.mcgill.ca
McGill University, Canada

Andrew Janco
ajanco@haverford.edu
Haverford College, United States of America

Quinn Dombrowski
qad@stanford.edu
Stanford University, United States of America

Keywords: Russian, NLP, working groups
Topics: Global, English, Contemporary, digital research infrastructures development and analysis, Literary studies, Central/Eastern European Studies

Scholars who apply digital humanities tools and methods to non-English languages face a unique set of challenges, particularly when the language in question is not a national language in the country where they work. Most technical tutorials for DH leave the expected language unstated, with an implicit assumption that readers will be working with English. Non-English, non-national language literature departments at institutions worldwide tend to be small, making it more likely that anyone using DH methods in these departments will have to do so without a local support network of

colleagues who are engaged in working through the issues that inevitably arise.

The Russian Natural Language Processing working group (Russian NLP) models a new approach to addressing the resource and support needs of a distributed network of scholars who apply DH methods to a non-English, non-national language. Formed in 2019 with support from Stanford University's Division of Literatures, Cultures, and Languages, the Russian NLP working group has brought together graduate students, faculty, librarians, and staff who work with Russian literary and/or historical texts. The group meets monthly, despite the scheduling challenges inherent in the wide dispersal of its membership, ranging from the west coast of the United States to Moscow. During the course of the 2019-2020 academic year, the group will identify natural-language processing tools and libraries available for working with Russian, and will find or build corpora that are representative of the materials that underpin the participants' own research areas (e.g. 20th century diaries and letters, 19th century novels and plays, 21st century internet text). The group will evaluate the existing tools using these corpora, and will write up reviews of the tools' performance when applied to different kinds of texts, in order to facilitate other scholars' decision-making about which tools to adopt for their own DH work with Russian-language materials. In addition, the group will be writing technical tutorials, and is exploring the possibility of collaborating with the developers of English-language NLP tools to adapt them for use with Russian.

This poster will present the major results of the group's work for this year (e.g. the evaluation of existing tools for different kinds of Russian corpora), as well as reflections on the group's organization and approach, and the challenges it has encountered. We anticipate that this will be of interest and value to anyone working with non-English materials who wants to cultivate a virtual community of practice around applying DH methods to their language.

Thinking Backwards, Walking Ahead: Guiding Humanities Researchers through Digital Project Life Cycles

Caitlin Pollock
pollock.caitlin@gmail.com
University of Michigan, United States of America

Joe Bauer
joebauer@umich.edu
University of Michigan, United States of America

Keywords: workshop, project management, data management, human-centered project design, backwards modeling
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, project design, organization, management, Education/ pedagogy, Library & information science

In order to avoid outdated projects, unpreserved knowledge, uncredited labor, and privacy or consent issues with digital projects University of Michigan has started providing a series of workshops that emphasize a change in culture and process in the project life cycle. Workshop participants learn how to visualize the life cycle of a project using human-centered design and backwards modelling when planning their projects to better understand how to version, archive, and preserve their research projects. This poster shares our experience with designing and running this workshop series at the University of Michigan and will give us an opportunity to connect with peers to build a community of practice for digital humanities scholars who develop and lead workshops. Materials from the workshop series will be shared as open data.

Assessing the Impact of an Undergraduate Digital Humanities Summer Research Program

Courtney M. Paddick
cpaddick@bloomu.edu
Bloomsburg University, United States of America

Carrie Pirmann
cmp016@bucknell.edu
Bucknell University, United States of America

Keywords: assessment, undergraduate research, pedagogy

Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, Library & information science

In 2017, librarians at Bucknell University developed a librarian-led undergraduate digital scholarship research program. We created the Digital Scholarship Summer Research Fellows (DSSRF) program to broaden research opportunities for students and introduce them to new ways of engaging in scholarship. We drew on our previous experience working on digital scholarship and digital humanities projects at Bucknell, through which we had developed skills in GIS, text analysis, and data visualization, as well as our backgrounds as research and instruction librarians to inform the creation of DSSRF. The program operates on a cohort model, with students meeting as a group throughout the summer under the guidance of two librarians. The eight week program provides students with an opportunity to undertake independent research on a topic of their own choosing, and utilize digital humanities tools and methodologies to both answer questions and convey their research findings.

Thus far, there has been limited research on the impact of digital humanities summer research or internship programs. Existing research on the impact of undergraduate research programs focuses on specific fields, e.g., the sciences (Lopatto, 2010). Through DSSRF, we have integrated informal assessment, both throughout the summer and at the conclusion of the program; with the 2019 cohort, we also conducted a more formalized assessment survey. However, these assessments only speak to the impact of the program in the moment, and do not provide any data on how the DSSRF experience impacts students in the months and years after they complete the program.

We surveyed past Digital Scholarship Summer Research Fellows to understand how their participation and the skills they acquired were applicable to their subsequent coursework and career paths, and how the program influenced their thinking about scholarship.

Some of what we hope to learn from our survey includes:

- If and how students put skills or tools learned in DSSRF to use in other contexts
- Did participation in DSSRF influence future course selections, majors/minors, or their career trajectory?
- What research, presentation/public speaking, writing, and soft skills students developed by participating in an immersive research experience
- Changes in the students' understanding of digital humanities as a field
- How did working with digital scholarship and digital humanities influence their thinking about scholarship more broadly?

We plan to utilize what we learn through this assessment process to make changes to future iterations of the DSSRF program, with the goal of making the program a more holistic and student-centered experience. We hope that sharing this information with the broader digital humanities community will be informative to others facilitating or considering developing undergraduate summer research programs.

Bibliography

Bauer, Karen W. and Joan S. Bennett. "Alumni Perceptions Used to Assess Undergraduate Research Experience." The Journal of Higher Education vol 74, no. 2 (2003): 210-30.

Digital Scholarship Summer Research Fellows. Bucknell University, 2019, Accessed 21 October 2019. https://dssrf2019.blogs.bucknell.edu/

Lopatto, David. "Undergraduate Research as a High-Impact Student Experience." Peer Review vol 12, no. 2 (2010): 27-30.

McDorman, Todd. "Promoting undergraduate research in the humanities: Three collaborative approaches." CUR Quarterly 25 (2004): 39-42.

Morris, Sarah. "The Digital Humanities Summer Scholarship: A model for library-led undergraduatedigitalscholarship."College & Undergraduate Libraries 24, no.2-4? (2017):532-544.

Wertzberger, Janelle and R. C. Miessler. "Dreaming Big: Library-led Digital Scholarship for Undergraduates at a Small Institution." In Undergraduate Research and the Academic Librarian: Case Studies & Best Practices, edited by Stephanie Davis-Kahl and Merinda. K. Hensle y,83-92. Chicago: The Association of College & Researc h Libraries, 2017.

Noh as Intermedia: Interactive Multi-Scale, Multimedia Analyses of Full-Length Noh Theater Performances

Takanori Fujita
Kyoto City University of Arts, Japan

Jaroslaw Kapuscinski
Stanford University, United States of America

François Rose
Stanford University, United States of America

Scott Bailey
Stanford University, United States of America

Peter Broadwell
broadwell@stanford.edu
Stanford University, United States of America

Simon Wiles
simon.wiles@stanford.edu
Stanford University, United States of America

Glen Worthey
gworthey@stanford.edu
Stanford University, United States of America

Keywords: intermedia, noh theater, interactive, multimedia
Topics: Asia, English, 5th-14th Century, 15th-17th Century, 18th Century, mixed-media analysis, music and sound digitization, encoding, and analysis, Musicology, Performance Studies: Dance, Theatre

The Noh as Intermedia project uses web-based technologies for interactive annotation and visualization of performing arts materials to investigate and present how Japanese Noh theater exemplifies intermedia. "Intermedia" here is defined as a way of forming expression that draws on relationships between art forms (literature, vocal and instrumental music, stage movement, dance, masks, costumes, and props, in the case of Noh) and their cumulative impact.

To untangle some of the unique "coming together" of artistic media in Noh, the project site builds a multi-scale, "thick" analysis around high-resolution video recordings of live performances of two exemplary and contrasting Noh plays, Hashitomi and Kokaji, which were both likely written in the late 14th century CE. The examination moves from high-level essays describing the genre and history of Noh and the two plays to a middle-level set of interactive catalogs, descriptions, and image, video and audio examples presenting the varied and often highly systematized media elements that comprise a Noh performance. These include theatrical and dramatic forms at the macro and meso scales, poetic modes of declamation and singing, dance patterns and other gestures, instruments, rhythmic and melodic modes and musical forms, as well as costumes, masks, stage positioning and props.

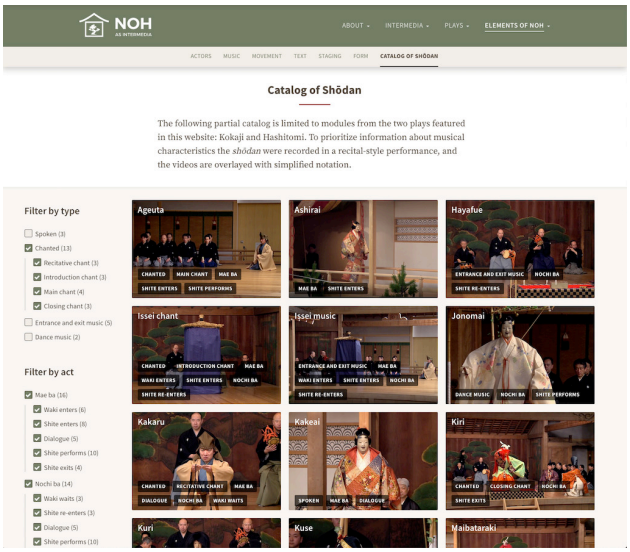


Fig. 1: A faceted catalog of shodan (formal sections of a play)

At the most finely detailed level, the project provides a novel playback and annotation environment that is synchronized at subsecond-level granularity with the streaming video playback, combining plot synopses, visualizations of formal sections, and synchronized libretto translations with a multilayered, multimedia score that encodes every sung or spoken word, musical element, and stage action. This score thus represents, for a specific performance, even the components of a play that are not traditionally notated, such as elements of the musical accompaniment. In an analogous manner, this entire mode of presentation has the effect of exposing, in an immediately engaging and comprehensible way, the myriad subtle details of a Noh performance whose interplay across media types lends the performances their extraordinary richness, yet which would otherwise likely go unremarked by those who are not already experts in the genre.

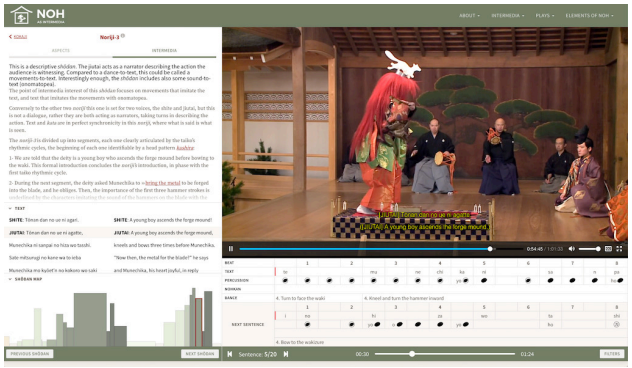


Fig. 2: The interactive playback environment, with synchronized score, text, formal map and description

The site also engages with the tensions between providing interactive, multimedia visualization and annotation features that are highly tailored to a specific performance genre (Noh, in this case) versus having the flexibility to accommodate other genres and art forms within the same framework. In our project, we found that favoring the specific over the general provided the greatest degree of interpretive insight, but it consequently required a greater effort to integrate other genres, particularly if these other forms are to be presented at equivalent levels of coherence and clarity.

References

Bethe, Monica, and Karen Brazell. Dance in the No Theater. Ithaca, N.Y.: China-Japan Program, Cornell University, 1982.

Chion, Michel. L'audio-Vision: Son et Image Au Cinéma. Nathan, 1990. Later translated as Chion, Michel, and Claudia Gorbman. Audio-Vision: Sound on Screen. Columbia University Press, 1994.

Cook, Nicholas. Analysing Musical Multimedia. Clarendon Press, 1998.

Hare, Thomas B. Zeami's Style: The Noh Plays of Zeami Motokiyo. Stanford, Calif: Stanford University Press, 1986.

Higgins, Dick with an appendix by Hannah Higgins. "Synesthesia and Intersenses: Intermedia." Something Else Newsletter 1. Something Else Press, 1966.

Konparu, Kunio. The Noh Theater: Principles and Perspectives. Warren, Conn: Floating World, 2005.

Tamba, Akira. The Musical Structure of No. Tokyo: Tokai University Press, 1981. Originally Tamba, Akira. La Structure Musicale Du No: Théâtre Traditionnel Japonais. Paris: Klincksieck, 1974.

The Public Approach to Re-represent Chinese Rural Memory in Cyberspace a Case Study on Gaoqian Village

Linqing Ma
malinqing2010@126.com
Renmin University of China, People's Republic of China

Jihong Liang
Renmin University of China, People's Republic of China

Huiling Feng
Renmin University of China, People's Republic of China

Keywords: Rural memory, Digital arrangement, Digital re-representation, Cultural analysis
Topics: Asia, English, 19th Century, 20th Century, Contemporary, cultural analytics, digital archiving, Cultural studies, Informatics

Re-representing rural memory in cyberspace complying with the principle of publicity is an issue of realistic significance at the moment in China. This paper represents the research findings of a pilot project on "Digital protection and inheritance of historical and cultural villages and towns", taking Gaoqian village as a case. The main work carried out includes digital arrangement and re-representation of culture heritage of Gaoqian based on cultural analysis, which is the core of the project, and also where publicity basically relies on. The publicity of the approach manifests as followings: 1) It aims at constructing the rural cultural memory of the public; 2) The final products, including a digital repository of Gaoqian cultural materials and a website on Gaoqian rural cultural memory, are for the public; 3) The interested parties i.e. scholars, villagers, local government, and technology companies, work in close corporation.

Creating a New Semantic Model for Ancient Greco-Roman Prosopography- Toward a Contextual & Historical Description of the Prosopographical Data-

Jun Ogawa
htjk6513khbk@yahoo.co.jp
University of Tokyo, Japan

Satoru Nakamura
University of Tokyo, Japan

Ikki Ohmukai
University of Tokyo, Japan

Kiyonori Nagasaki
International Institute for Digital Humanities, Japan

Keywords: Greco-Roman study, prosopography, contextual description, semantic model, ontology
Topics: Europe, English, BCE-4th Century, data modeling, linked (open) data, History, Humanities computing

We will propose a way of describing historical prosopography considering carefully the chronological context.

Even though there are already several ontologies or vocabularies designed for historical prosopographical study, or even specifically for ancient study, most of them is not sufficient for dealing with the chronological change of the status of each person.

We therefore introduce a new model. In addition to common class entity 'Person', we create some unique class-entities 'Person@moment'. Since this entity indicate a temporal aspect of a person, we are now able to describe him not only as a single entity, but also as a gathering of contextual entities that each of them should be identified. And, same model can be applied to other objects like community(organization).

Thanks to this model, we are capable of describing 'Person' in the chronological context in more detailed way, and it surely is widely applicable in other fileds of historical study.

THE IMPORTANCE OF THE PRESERVATION OF HUMAN INVOLVEMENT OF SOUTH AFRICANS IN THE ANTARCTIC REGION AS THE ONLY COUNTRY IN AFRICA.

Maria Olivier
riaolivier@sun.ac.za
Stellenbosch University, South Africa

Keywords: Antarfctica, Marion Island, Gough Island, Humanities, South Africa
Topics: Africa, Global, English, 18th Century, 19th Century, 20th Century, data, object, and artefact preservation, digital archiving, History, Library & information science

The presentation will focus on the preserving of material linked to humanities of South Africa's involvement in the Antarctic region through a digital repository for the future generations of South Africa. The presentation will show the growth of the repository in the past decade through the collection of material through a small group of people involved in a specific geographical area. It will highlight that currently it is the only country within the Africa continent that is involved in the Antarctic region, therefore will highlight the importance of creating a open accessed digital repository.

Paper Feature Extraction from Digital Images

Shouji Sakamoto
sakamoto@mac.com
Ryukoku University, Japan

Keywords: paper analysis
Topics: Asia, Europe, English, 5th-14th Century, 15th-17th Century, 18th Century, image processing and analysis, Art history, Asian studies

Digital archives are very useful to readers. However, they do not give enough information, such as paper, ink, and pigment, to satisfy researchers of materials and conservators of the documents. In this research, we describe paper feature extraction, such as laid and chain line pattern, watermark, syanomepattern, and itamepattern from digital images of paper.

Ten years recovering the memory of republican exile with citizen collaboration. The results of E-xiliad@s Project a perspective from the Digital Humanities and the Digital Public History.

Lidia Bocanegra Barbecho
lbocanegra@go.ugr.es
Universidad de Granada, Spain

Keywords: Crowdsourcing, Citizen science, Spanish Republican exile, Digital Public History, Open Data
Topics: South America, Europe, English, North America, 20th Century, Contemporary, crowdsourcing, public humanities collaborations and methods, History, Humanities computing

The Spanish republican exile was the result of the Republican defeat in 1939 by the Francoist army, led by the general Francisco Franco. Nearly half a million-people had to go into mass exile during the months of January and February, through the French border crossings. Many other exiles did so, months later, from Alicante to the North African coasts. These places of destination were, in most cases, places of passage to successive destination countries in Europe and, especially, in Latin America. The international nature of this historical event means that there is currently a large number of personal files scattered in different places around the world. In order to recover these stories, the e-xiliad@s project was conceived in 2009, with a Digital Humanities and Public History perspective: www.exiliadosrepublicanos.info. It is an crowdsourcing project that, through a multilanguage digital platform, retrieves unpublished documents about the anonymous exiled. From the research point of view, the privileged target audience is composed by relatives and friends of the exiles and those interested in the subject. This initiative funded twice (2009 and 2011) by the General Directorate of Migrations of the Spanish Ministry of Employment and Social Security, uses a methodology created ad hoc to obtaining data based on public participation from citizen science. That is, the content is generated on-line by the public at an international level and coordinated by a scientific specialist. For almost a decade, this crowdsourcing project has been developing an online public engagement strategy for public participation based on open data, supported by a custom digital platform and its digital social networks, with more than 1.500 followers. At this stage, the project recovered around five hundred unpublished archives among photographs, memories, official documents, letters and interviews, that comes associated with about two hundred completed exile records. The vast majority of these data are public, thanks to the informed consent of the author.

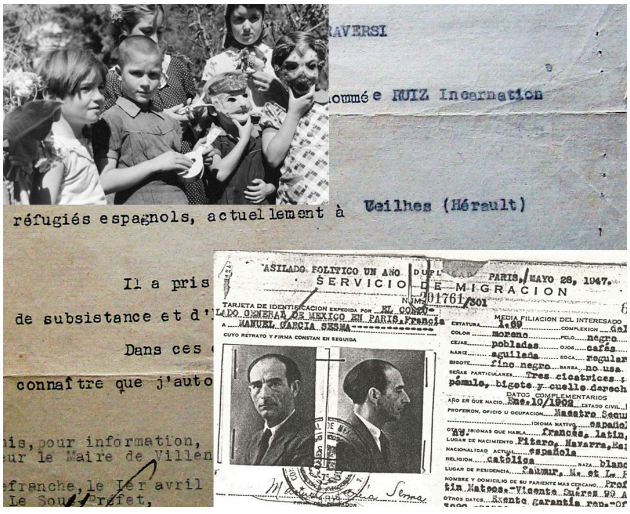


Figure 1. Examples of documents and photos provided by relatives of exiles registered on the project platform and collaborators.

At technological level, e-xiliad@s has been built using Drupal 6 LTS (Long Term Support) with a MySQL database. The technological solution answer the need of the internal survey form (figure 2) that has been

created using a specific data model connected with a strategic communication plan, that serves to get data online via user' confidence with the project and to stimulate his/her family memoir.

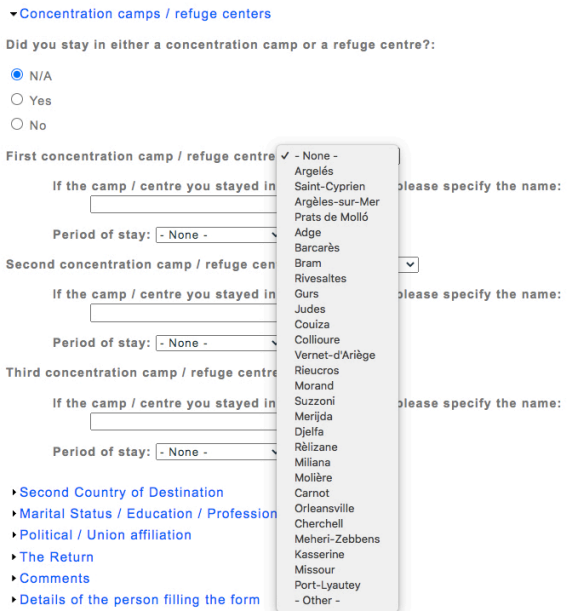


Figure 2. Detail of the internal form to be filled

As lesson learned, the need to maintain a technology that is getting older without official support (Drupal); the management of a project social network community continuously growing; and the spamming problem due to the project popularity within the republican exile community (more than 11.000 spamming users) are questions that we are trying to solve. Despite this, E-xiliad@s acts as a digital identity place for those connected to this topic and as a space that informs society, with scientific rigor, from the field of Digital Public History. Currently there are six large online social network communities that move much of the information on the subject of Republican exile and the Spanish Civil War, among which is the e-xiliad@s project with its social networks (Facebook: @exiliados.republicanos and Twitter: @exiliadas). They play a significant role at the national and international level regarding the recovery of historical memory, due to the quantity and quality of the information it offers (figure 3).

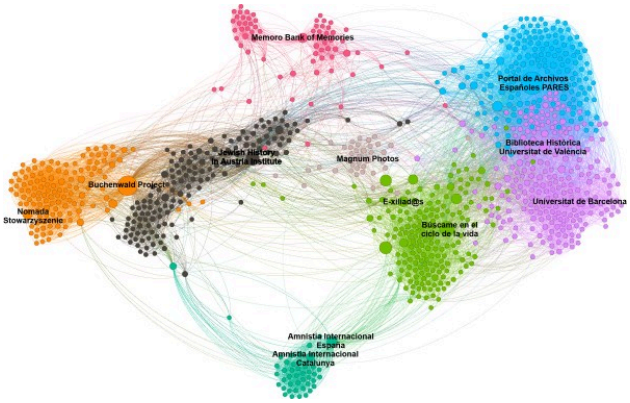


Figure 3.Graph analysis that represents the first and second level connections of the social networks of the E-xiliad@s project, which is represented in the community in green.

CREDITS:

This contribution has been funded by the research project, with title: Análisis de la Participación Pública en la investigación histórica desde el ámbito de la ciencia ciudadana (Co-Historia), funded by the Programa Operativo FEDER 2018, Universidad de Granada, Ref. E-HUM-507-UGR18. Principal Investigator: Lidia Bocanegra Barbecho.

BIBLIOGRAPHY:

Bocanegra Barbecho, Lidia, & Toscano, Maurizio. (2015). El exilio republicano español: Estudio y recuperación de la memoria a través de la web 2.0. Nuevo enfoque metodológico con el proyecto e-xiliad@s. Migraciones & exilios, 15, 113–136. http://doi.org/10.5281/zenodo.1182238.

Bocanegra Barbecho, Lidia; Toscano, Maurizio; Delgado Anés, Lara. (2017). Co-creación, participación y redes sociales para hacer historia. Ciencia con y para la sociedad. Historia y comunicación social, 22(2), 325–346. http://doi.org/10.5209/HICS.57847.

Bocanegra Barbecho, Lidia, & Toscano, Maurizio. (2016). The Spanish Republican Exile: Identity, Belonging and Memory in the Digital World. In Cultural Heritage in a Changing World (pp. 237–253). Springer International Publishing. http://doi.org/10.5281/zenodo.1182426.

E-XILIAD@S SOCIAL NETWORKS:

@exiliados.republicanos: https://www.facebook.com/exiliados.republicanos

@exiliadas: https://twitter.com/exiliadas

Introducing an Automated Pipeline for a Browser-based, City-scale Mobile 4D VR Application Based on Historical Images

Sander Muenster
sander.muenster@uni-jena.de
FSU Jena, Germany

Ferdinand Maiwald
TU Dresden, Germany

Mathias Hofmann
TU Dresden, Germany

Walpola Perera
TU Dresden, Germany

Keywords: pipeline, 3Dcity model, browser based viewer
Topics: EuropeEnglish19th Century20th Century, Contemporary, virtual and augmented reality creation, systems, and analysis, Computer science, History

The process for automatically creating 3D city models from contemporary photographs and visualizing them on mobile devices is now well established, but historical 4D city models are more challenging. The fourth dimension here is time. This contribution describes an automated VR pipeline based on historical photographs and resulting in an interactive browser-based device-rendered 4D visualization and information system for mobile devices. Since the pipeline shown is currently still under development, initial results for stages of the process will be shown and assessed for feasibility.

Reconsidering the Roman Workshop using machine learning to examine the processes behind inscribed texts

Charlotte Tupman
c.tupman2@exeter.ac.uk
University of Exeter, United Kingdom

Jacqueline Christmas
University of Exeter, United Kingdom

Dmitry Kangin
University of Exeter, United Kingdom

Keywords: machine learning, computer science, neural networks, epigraphy, digital classics
Topics: Europe, English, BCE-4th Century, artificial intelligence and machine learning, image processing and analysis, Computer science, History

Introduction

This poster addresses a collaboration between ancient historians and computer scientists which uses machine learning to analyse the planning processes involved in creating Latin inscriptions. It represents a pilot project funded by an Institute for Data Science and Artificial Intelligence Research Award at the University of Exeter, led by Dr. Charlotte Tupman, Dr. Jacq Christmas and Dr. Dmitry Kangin.

This contribution addresses the theme of the open data movement: its entire dataset comprises materials made available by the Epigraphische Datenbank Heidelberg (EDH) as open data. This project could not have taken place without the principles of open data being embraced by the digital classicist community and the EDH team.

Aims

Inscriptions are one of our most important sources for the ancient world, covering a vast number of subjects, monument types, geographical areas and periods [2]. There were several stages in their planning: the preparation of the surface, drafting of the layout of the text, possibly the filling in of letters with brush strokes, and finally the carving [3, 4]. Our interest is in

the drafting: were Roman craftsmen in different regions working to defined sets of planning 'modules' which could be scaled up or down? Did the stonecutter have freedom to influence the design after drafting? Does this vary by time or place? Could we even begin to identify the work of specific workshops [5]?

Methodology

The project applies a neural network, investigating Neural ODEs [6] and extending the model for this image processing case. We investigate whether the model can learn effectively, and address the development of a new basis for training Neural ODEs using a mathematical method, invariant embedding. Once characters in an image have been localised, we analyse the regularity of their size, shape, spacing, position and orientation, and the overall shape of their outline.

Summary

Improving our understanding of the planning out of inscriptions is important not only from the perspective of ancient craftsmanship. Inscriptions are frequently fragmentary, and epigraphers have to determine what letters would have been present on the missing sections. Reconstructions of inscriptions influence our interpretation of ancient sites, so being able to predict accurately the positioning of the letters is crucial. We seek to initiate valuable conversations with DH specialists in other fields which may influence the future direction of our investigations.

References

[1] Epigraphic Database Heidelberg, https://edh-www.adw.uni-heidelberg.de/

[2] J. Bodel (ed.), 2001. Epigraphic Evidence: Ancient History from Inscriptions, London, p. 4

[3] R. Grasby, 2009. Processes in the Making of Roman Inscriptions. Study 11, p.12

[4] G. Susini, 1965. The Roman Stonecutter, Oxford

[5] B. Russell, 2013. The Economics of the Roman Stone Trade, Oxford

[6] T.Q. Chen, Y. Rubanova, J. Bettencourt and D.K. Duvenaud, 2018. Neural Ordinary Differential Equations. In Advances in Neural Information Processing Systems, pp. 6571-6583

Toward Exploring Artist Information from Seal Images in Ukiyo-e Collections

Kangying Li
gr0319ss@ed.ritsumei.ac.jp
Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

Biligsaikhan Batjargal
Kinugasa Research Organization, Ritsumeikan University, Japan

Akira Maeda
College of Information Science and Engineering, Ritsumeikan University, Japan

Ryo Akama
College of Letters, Ritsumeikan University, Japan

Keywords: image retrieval, ancient documents processing, retrieval-based recognition
Topics: Asia, English, 18th Century, 19th Century, 20th Century, image processing and analysis, information retrieval and querying algorithms and methods, Computer science, Library & information science

Ukiyo-e served as ambassadors bringing Japanese culture to the world, and also Ukiyo-e artists were expected to anticipate trends and always stayed a step ahead of new fashion and style .We could know who is the artist of a certain print from signatures or seals in Ukiyo-e painting. A system that automatically recognizes signatures and seals can help enthusiasts and professionals better understand the background information of these collections more efficiently. In this research,we focus on automatic identifying of signatures and seals in Ukiyo-e painting,and we use open data to provide scholars a new perspective to Ukiyo-e collections.

Information extraction across textual corpora: semi-automatic text-tagging workflow with Chinese local gazetteers

Calvin Yeh
cyeh@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Sean Wang
swang@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Shih-Pei Chen
Max Planck Institute for the History of Science

Keywords: text-tagging, information extraction, local gazetteers, Chinese history
Topics: Asia, Comparative (2 or more geographical areas), Global, English, 15th-17th Century, 18th Century, 19th Century, text encoding and markup language creation, deployment, and analysis, text mining and analysis, History of science, Humanities computing

Textual information extraction is necessary for many humanities projects. Since 2013, we have been developing "Local Gazetteers Research Tools" (LoGaRT), and its text-tagging component is designed for that purpose. This poster introduces the practical implementation of information extraction and organization in LoGaRT and discusses how this component could be applied to other corpora with consistent internal structures.

Stylo, a semantic writing tool for scientific publishing in Human Sciences

Margot Mellet
margot.mellet@umontreal.ca
Canada Research Chair on digital textualities

Antoine Fauchié
antoine.fauchie@umontreal.ca
Canada Research Chair on digital textualities

Nicolas Sauret
Canada Research Chair on digital textualities

Marcello Vitali-Rosati
Canada Research Chair on digital textualities

Arthur Juchereau
Canada Research Chair on digital textualities

Keywords: Content production, Publishing, Digital platforms, open access, Publishing platforms
Topics: Europe, English, North America, Contemporary, digital publishing projects, systems, and methods, scholarly editing and editions development, analysis, and methods, Humanities computing

Stylo (https://stylo.huma-num.fr/articles) is a tool designed by the Canada Research Chair on Digital Textualities, with the support of Érudit, designed to simplify writing and editing scientific articles in the humanities and social sciences (HS).Writing in a digital environment whether for printed or digital dissemination is now a matter-of-course: no one writes an article by hand and then has it typed. However, rarely do scholars think about issues related to the act of writing and, in addition, we have far from exhausted the potential for digital technologies to enhance writing and editing practices. Moreover, the tools available for writing in an academic context are limited, often proprietary and rarely designed and adapted to the needs of HS researchers. In concrete terms, almost all scientific production in HS today uses a single software and the same format: Word and docx, both of which are owned by Microsoft. Designed for all contexts although particularly intended for office use, as the name indicates, this duo is poorly adapted to the specific needs of scholars. Stylo is also a tool for writing and editing scientific texts, specifically designed to meet the challenges of scholarly digital publishing. Stylo is still a prototype; its development remains strongly linked to theoretical thinking on writing in a digital environment. Many less successful software have and continue to be developed. However none have come close to rivalling Microsoft's dominance on the world stage of writing tools. The ambition of the project is not so much to create an umpteenth tool, but to think about the epistemological stakes of writing practices in digital environments. How do these new forms of writing structure thought? The philosophy of Stylo builds on the simple fact that writing skills are semantic rather than graphic, and yet, authors of scholarly articles are often asked to deliver graphical layout rather than enriched text. Indeed, writing tools have a major impact on the production of knowledge: they influence content structure; define content accessibility criteria; determine the durability or obsolescence of content.Almost all scientific content is still produced with proprietary word processing software that is not conceptualised and designed for

academic use, namely Microsoft Word (Kirschenbaum Track Changes: A Literary History of Word Processing 2016). However, this tool for the general public has problematic shortcomings for scientific publishing in HS:- The principle of WYSIWYG (What You See Is What You Get) confuses structuring and formatting: to facilitate the reading and processing of an article or book, the semantic value of fragments must be dissociated from their graphic rendering, for example: a level 2 title and its graphic rendering in a large font size. Stylo belongs to this type of initiative, alongside Software Studies (Fuller Software Studies: A Lexicon 2008).- The docx format does not allow the contents page to be structured scientifically. The contents produced in docx are poorly tagged and difficult to access because they are not indexed. At a time when text production has never been so abundant, it is crucial that documents be properly structured to improve their indexing and allow them to be queried using advanced search tools.- Publishers undertake the long and tedious task of reworking texts in order to reintroduce a semantic layer and to correct the many errors created by these word processors.- The durability of documents in docx format depends on Microsoft's willingness to maintain this proprietary format. However, sustainability is neither the mission nor the priority of a commercial company. There is no guarantee that content in docx format will be accessible in the future.By closely studying the editorial practices of scientific journals in HS, it appears that, when passing from the author to the publisher to the distributor, the scientific article goes through highly disparate states and formats. This process results in a considerable loss of time, data and meaning.There is therefore an urgent need to conceptualise and experiment with new writing models that are better adapted to the needs of HS knowledge production.Stylo has been conceived to offer a complete and continuous workflow from writing to publishing and to avoid loss of data and meaning between the various editing stages. It gathers functionalities such as sharing, versioning, track changes, reference management, revision annotations, multi-format export, metadata aligned with online authorities (LOC, Rameau, Wikidata, ORCID, ...) and inline semantic markup. In order to avoid frequent data breaches during editing processes, Stylo offers a continuous workflow from writing to dissemination, ensuring that semantic information provided by the author will indeed be published.Stylo can provide multiple outputs, depending on the needs of the author and the publisher:1. XML (Érudit, JATS, TEI, etc.) for diffusion and harvesting platforms2. HTML5 for direct publication on CMS3. PDF & EPUB with programmable stylesheets4. ODT, DOCX for common useBased on modular, lowtech and standard editing tools and formats, such as markdown, bibtex, Pandoc, Hypothes.is and LaTeX, Stylo integrates, into a single interface, the best practices in terms of web-based writing and publishing.In the digital chain of writing and publishing scholarly journals in HS, it is increasingly clear that the publisher no longer has control over the cultural, social or technical existence of its contents in an environment that is increasingly dependent on text-related data. Indeed, making a text exist in the digital environment implies exposing sufficiently rich and relevant data so that the contents be indexed by search engines. Stylo contends that each actor, from the author to the disseminating platform, is likely to be involved in a production process that guarantees the continuity of data. This is possible thanks to the sharing of a single pivotal format that is simple to access and use, and rich enough for the publishing of scientific texts in the humanities and social sciences.In this vision, writing is no longer just writing. Writing in the digital environment becomes part of a publishing dynamic. In other words, the act of writing necessarily goes hand in hand with the act of publishing. This idea that writing = writing + structuring brings us closer to a culture of digital literacy, which enables a mastery over the writing environment and an ability to inscribe space as much as to structure it. It is a matter of both knowing how to evolve in this environment and how to make it evolve.As pragmatic research project, Stylo works alongside the practitioners of writing and scholarly publishing, aiming for an evolution and an enrichment of these practices and inviting actors to reflect on the most effective means to harness the power of the digital environment.

Les Ateliers de [sens public] de l'édition collaborative au livre ouvert

Hélène Beauchef
Canada Research Chair on digital textualities

Antoine Fauchié
antoine.fauchie@umontreal.ca
Canada Research Chair on digital textualities

Servanne Monjour

Canada Research Chair on digital textualities

Nicolas Sauret
Canada Research Chair on digital textualities

Marcello Vitali-Rosati
Canada Research Chair on digital textualities

Michael Eberle-Sinatra
Canada Research Chair on digital textualities

Keywords: digital publishing, academic monograph, publishing model, digital textualities, publishing chain
Topics: French, North America, Contemporary, digital publishing projects, systems, and methods, scholarly editing and editions development, analysis, and methods, Book and print history, Literary studies

Quel est le futur du livre ? Est-ce que les livres papier existeront encore dans quelques années ? Ou sera-t-on passé au tout numérique ? Que peut apporter le support numérique dans le domaine de l'édition universitaire ? Ce sont des questions qui se posent encore avec urgence aujourd'hui, à l'heure où la publication académique a certes fait des progrès en matière de diffusion (plateformes de diffusion, libre accès), mais encore très peu d'avancées sur le plan éditorial. En témoigne par exemple la prédominance du format PDF pour les versions papier, ou encore les limites du format EPUB. Ces deux formats, destinés spécifiquement à une édition numérique homothétique au papier, ne permettent pas à la monographie d'embrasser pleinement les potentialités de l'environnement numérique. Une telle ambition suppose que les acteurs en Sciences humaines (SH) s'emparent de leur propre futur en le construisant.Au carrefour des pratiques de lecture classiques et des nouveaux usages liés au numérique, la collection augmentée des Ateliers de [sens public] (accessible à l'adresse https://ateliers.sens-public.org/) propose une complémentarité entre édition papier et édition numérique. Conçus par des chercheures en SHS, les ouvrages publiés par les Ateliers se distinguent de la monographie académique traditionnelle en explorant des formes d'écriture alternatives : essai, manifeste, échanges épistolaires, carnet de recherche, etc. L'objectif de la collection est de proposer des modèles d'écriture et de publication favorisant la conversation entre les chercheures, la réappropriation du savoir et l'éditorialisation des contenus publiés. Un tel modèle suppose la mise au point d'une chaîne éditoriale innovante qui engage une attention différente, disséminée et distribuée, privilégiant l'approfondissement par rebonds.Les deux formes de publication qui fondent la collection (édition papier et édition numérique) présupposent des pratiques différentes de lecture et deux approches différentes à la réception des contenus.Le livre papier - ainsi que le numérique homothétique (EPUB ou PDF) qui reproduit à l'identique le livre papier sur un support numérique - tend à une lecture linéaire. Une thèse peut y être présentée et argumentée de façon complexe. Le lecteur sera capable de suivre de manière linéaire le développement de l'argumentation, de reprendre le cheminement de pensée de l'auteur en se laissant accompagner d'un bout à l'autre du discours. La collection impose ainsi un format assez court (120/200 pages), à savoir la longueur adéquate pour présenter une thèse et la démontrer à l'aide d'une argumentation unique et cohérente. La lecture savante d'un livre papier suppose de passer plusieurs heures pour saisir tout le raisonnement de l'auteur. Le discours doit alors nécessairement être linéaire, amenant le raisonnement à sa fin. On cherche alors à minimiser les écarts par rapport au chemin principal. L'appareil critique, les références, les parenthèses, les exemples, les images, les statistiques, les détails... tous ces éléments participent à altérer la linéarité du discours et de la lecture.L'édition numérique augmentée, en revanche, présuppose une lecture non linéaire, qui procède par approfondissement et par associations. L'expérience de lecture sur un sujet débute avec un premier texte dont on souhaite approfondir un aspect. Ce qui semblait n'être qu'un détail se révèle foisonnant lorsque l'on navigue vers un contenu associé mais localisé ailleurs. Cette navigation génère le parcours emprunté par le lecteur. Ce parcours n'a pas été prévu par un auteur ou un éditeur, et ne pré-existe pas à la navigation. Si l'auteur ou l'éditeur ont suggéré des pistes et ouvert des portes vers d'autres contenus, la navigation est effectivement laissée aux lecteurs, selon leurs envies, leurs besoins ou encore leur créativité, c'est-à-dire leur capacité d'association. En ce sens, il n'est pas exact, comme le voudrait Nicholas Carr, que le numérique détruit notre capacité d'attention. En matière de lecture savante, le support numérique de lecture ne favorise pas le suivi d'un discours long et unitaire, mais promeut une attention différente, disséminée, privilégiant l'approfondissement par rebonds. Une telle lecture suppose des efforts d'association qui relèvent déjà d'une activité herméneutique. En libérant le parcours de lecture, c'est une nouvelle responsabilité d'appropriation et d'interprétation qui est transférée dans les mains du lecteur. De son côté, l'éditeur se voit déléguer une nouvelle fonction éditoriale : celle

de concevoir et de mettre en place les conditions de l'appropriation et de l'interprétation.Deux objectifs principaux ont motivé le projet des Ateliers de [sens public] : Réfléchir aux modèles éditoriaux de l'édition numérique enrichie des monographies pour redéfinir un modèle épistémologique pour la monographie - structuration des contenus, balisage, rapports entre papier et numérique - qui réponde aux défis actuels de l'édition numérique. Fournir une solution logicielle adéquate et opérationnelle pour pouvoir produire des monographies enrichies.La collection des Ateliers de [sens public] offre les deux possibilités de lecture : une édition papier (avec ses analogues homothétiques numériques : EPUB et PDF) et une édition en ligne augmentée. L'édition en ligne augmentée propose le texte en intégralité et en libre accès, ainsi que toute une série de contenus additionnels qui sont autant de portes à franchir pour le lecteur, que ce soit des approfondissements, des sujets connexes, d'autres formes de contenus, d'autres plateformes, d'autres parcours. Le parcours de lecture, propre à chaque lecteur, n'a pas été spécifiquement prévus par l'auteur ou par l'éditeur du livre. C'est également pour cette raison que l'édition en ligne augmentée est gratuite : elle devient un point d'accès vers un univers connecté externe au livre, qui lui échappe et qui n'a été créé ni par l'auteur, ni par l'éditeur. L'édition numérique facilite par ailleurs la navigation à l'intérieur du livre, en proposant des index et des renvois qui démultiplient encore les parcours de lecture. En plus d'une circulation libre des contenus, les liens vers d'autres contenus mettent en place un cercle vertueux entre le travail de l'auteur et celui d'autres auteurs. Ils matérialisent un réseau de connaissances et un dialogue d'idées, sans pour autant sacrifier le discours linéaire, dont la forme de l'argumentation demeure claire, identifiée et reconnaissable dans le champ scientifique. Ces parcours propre à l'édition numérique augmentée sont complémentaires d'une lecture plus classique se plongeant pleinement dans la thèse de l'auteur et dans son cheminement de pensée, propre à l'édition papier ou à ses doubles numériques homothétiques.

Références

Guédon, Jean-Claude, et Alain Loute. 2017. « L'histoire de la forme revue au prisme de l'histoire de la grande conversation scientifique . Entretien avec Jean-Claude Guédon réalisé par Alain Loute, préparé avec l'aide de Caroline Glorie, Thomas Franck et Andrea Cavazzini. » Cahiers du GRM. publiés par le Groupe de Recherches Matérialistes Association, n? 12 (décembre). https://doi.org/10.4000/grm.912.Kaplan, Frédéric. 2014. « La question de la langue à l'époque de Google ». In Digital Studies : organologie des savoirs et technologies de la connaissance, 189. Innovation. FYP éditions.Larivière, Vincent, Stefanie Haustein, et Philippe Mongeon. 2015. « The Oligopoly of Academic Publishers in the Digital Era ». PLOS ONE 10 (6):e0127502. https://doi.org/10.1371/journal.pone.0127502.Larivière, Vincent, et Delphine Lobet. 2018. « La mort des livres dans les sciences humaines et sociales, et en arts et lettres? » Acfas. Acfas. 14 juin 2018. https://www.acfas.ca/publications/decouvrir/2018/06/mort-livres-sciences-humaines-sociales-arts-lettres.Morozov, Evgeny. 2012. The Net Delusion: The Dark Side of Internet Freedom. Reprint edition. PublicAffairs.Morozov, Evgeny. 2013. To Save Everything, Click Here: The Folly of Technological Solutionism. First Trade Paper Edition edition. PublicAffairs.Sordi, Paolo, et Domenico Fiormonte. 2019. « Geopolitica Della Conoscenza Digitale. Dal Web Aperto All'impero Di GAFAM. » DigitCult -

Scientific Journal on Digital Cultures 4 (1):21?36. https://doi.org/10.4399/97888255263183.Vitali-Rosati, Marcello. 2014. « Les Revues Littéraires En Ligne : Entre Éditorialisation et Réseaux d'intelligences ». Études françaises 50 (3):83. https://doi.org/10.7202/1027191ar.Vitali-Rosati, Marcello. 2018. « À Quoi Servent Les Publications Scientifiques? » La vie de la recherche scientifique, n? 412:19?22. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/21016.Wormser, Gérard. 2018. Facebook. Les atelier de Sens public.

Visualizing Collocations in Religious Online Forums

Thomas Schmidt
thomas.schmidt@ur.de
University of Regensburg, Germany

Florian Kaindl
University of Regensburg, Germany

Christian Wolff
University of Regensburg, Germany

Keywords: Collocations, Distant Reading, Religious Studies, Reddit, Visualization
Topics: Europe, English, North America, Contemporary, social media analysis and methods, text mining and analysis, Sociology, Theology and religious studies

1. Introduction

One of the most influential concepts in Digital Humanities (DH) in recent years is Moretti's (2000) idea of Distant Reading, more precisely the application of computational methods to analyze and visualize large amounts of text to gather new insights. Distant Reading has led to various successful projects especially in literary studies and linguistics (cf. Jänicke et al., 2015) but also religious studies, e.g. to analyze famous religious texts (McDonald, 2014; Slingerland et al.; 2017; Verma, 2017). We want to build primarily upon the work of (Pfahler et al., 2018) who applied topic modeling on Muslim online forums to investigate what this community is predominantly talking about. They identified several main topic clusters about eating, family and politics which are talked about the most.

We want to further explore the application and potential benefit of Distant Reading-methods for the use case of religious online forums. Our research goal is to examine the content, language, topics and sentiments in religious online forums of different religious subgroups to identify differences and similarities and learn more about the way of life and beliefs of these communities.

While we explore multiple methods like named entity recognition, topic modeling and sentiment analysis, in the following contribution we report upon our results for the method of collocation analysis. Via collocations, we want to analyze differences in the way several religious key concepts are discussed in online forums of different religious subgroups.

2. Methods

We have chosen Reddit (https://www.reddit.com/) for data collection since it is rather easy to scrape and one of the largest platforms on the internet. Furthermore, various religious subgroups are represented enabling us to compare content more easily.

We have acquired all submissions (threads) for the time span of July 1, 2018 to July 1, 2019 for the three subreddits /r/Christianity (https://www.reddit.com/) , /r/Islam (https://www.reddit.com/r/islam/) and/r/Occult (https://www.reddit.com/r/occult/) . We chose the first two since they represent the two largest monotheistic religions and included the third one to also examine a rather esoteric religious direction.

We have acquired over 700,000 comments and around 50 million tokens (figure 1).

Metric/Forum	/r/Christianity	/r/Islam	/r/Occult	Sum
Submissions	28,896	4,123	8,275	41,394
Comments	618,719	64,886	76,387	759,992
Tokens	43,996,066	4,754,301	5,702,675	54,453,042
Sentences	2,897,575	300,854	365,962	3,564,391

Figure 1: Corpus statistics

We have chosen five as maximum length for a collocation and measure the strength of collocations via Pointwise Mutual Information (PIM) which scores the collocations based on their actual co-occurrence in the corpus in proportion to their expected co-occurrence if they were independent (Church & Hanks, 1989). To visualize collocations, we place the key concept in the middle and the collocations around them. The higher the PMI-value, the closer the concept. We also put the exact PMI-score on the edges.

3. Results

In the following we showcase the use case for the spiritual key terms "love", "religion" and "life" and highlight some insights we gained.

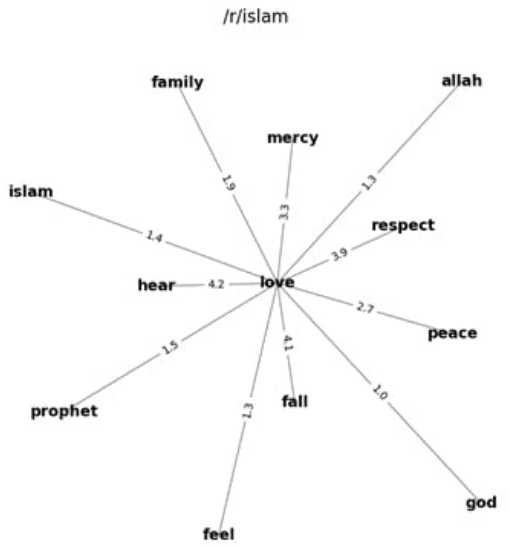


Figure 2: Collocations for "love" in /r/Islam

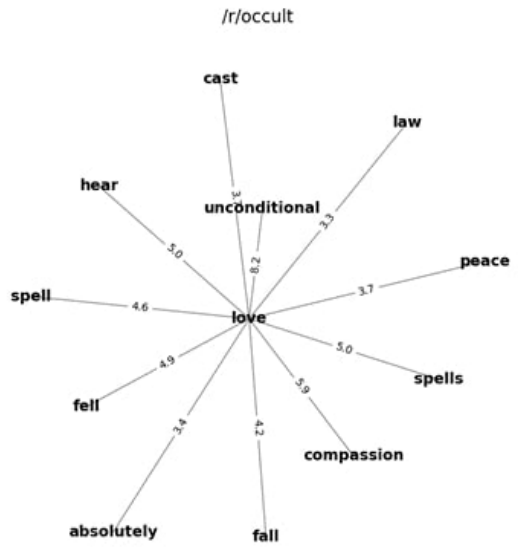


Figure 4: Collocations for "love" in /r/Occult

Figure 3: Collocations for "love" in /r/Christianity

In the Christian subreddit, we find that love shows most connections with idioms/quotes from the bible ("unconditionally", "enemies", "agape"; figure 3). In contrast, we find strong associations with positive terms, words for god and the prophet as well as for "family" in the Muslim forum (figure 2) which is in line with Pfahler et al. (2018) showing a strong focus on family-related topics in Muslim forums. For /r/occult we find rather fitting associations with the notion of magic, thus showing the rather esoteric content of this forum (figure 4).

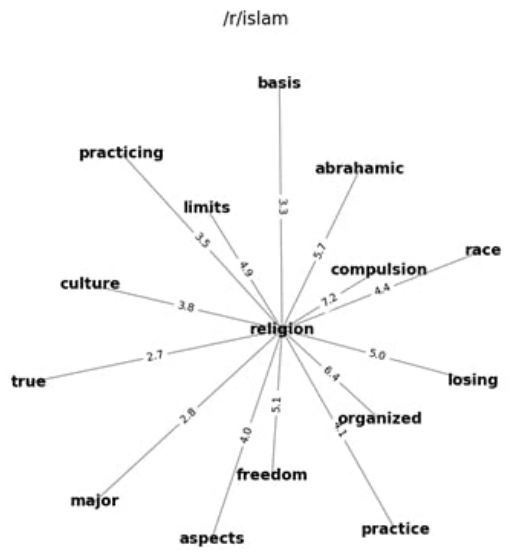


Figure 5: Collocations for "religion" in /r/Islam

Figure 6: Collocations for "religion" in /r/Christianity

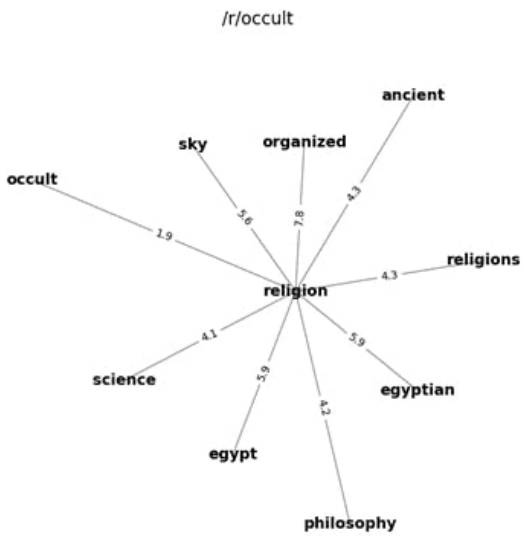


Figure 7: Collocations for "religion" in /r/Occult

Many terms in /r/Islam and the concept of religion point to discussions about religious directions e.g. "organized", "abrahamic", "culture", "major" (figure 5). The connection with race might be connected to the racism Muslims face in western countries. Quite similarly, /r/Christianity also shows collocations describing the discussion about other religions ("organized", "islam", "false") also pointing to rather heated discussions ("utter", "nonsense"; figure 6). /r/Occult shows collocations specifying the religion and other world views ("Egypt", "ancient", "philosophy", "science"; figure 7).

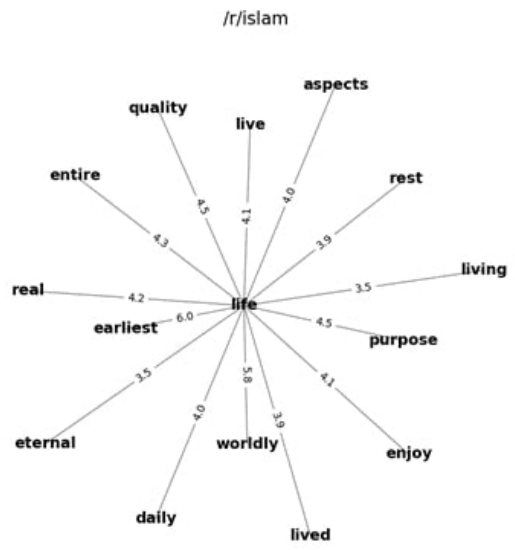


Figure 8: Collocations for "life" in /r/Islam

Figure 10: Collocations for "life" in /r/Occult

Figure 9: Collocations for "life" in /r/Christianity

In /r/Christianity, "life" is associated with words pointing to the afterlife ("everlasting", "eternal", "immortal") while in /r/Islam, it is rather tied to terms describing a direction in life ("purpose", "meaning"; figure 8 and 9). However, both subreddits show connections with rather positive words except for death concepts. Those collocations are indeed stronger for /r/Islam ("rest", "death", "short"). The collocations are quite varied for /r/occult (figure 10).

Overall, we were able to gather some first insights like the strong difference of /r/occult, connections to family and politics for some key concepts in the Muslim forum or the focus on discussions about religious directions for the concept of religion in all forums.

We plan to investigate other methods of computational text analysis but also want to apply more in-depth qualitative analysis of parts of our corpus via content analysis to confirm and evaluate some of our assumptions we derived via the collocation visualizations.

References

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. Proceedings of the 27th Annual Meeting on Association for Computational Linguistics, 76–83. https://doi.org/10.3115/981623.981633

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015, May). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In EuroVis (STARs) (pp. 83-103). McDonald,

D. (2014). A text mining analysis of religious texts. The Journal of Business Inquiry, 13(1), 27-47.

Moretti, F. (2000). Conjectures on world literature. New left review, 54-68.

Pfahler, L., Elwert, F., Tabti, S., Morik, K., & Krech, V. (2018). What do you do with 5 million posts?? Versuche zum distant reading religiöser Online-Foren. DHd Konferenz 2018, 335-338.

Slingerland, E., Nichols, R., Neilbo, K., & Logan, C. (2017). The distant reading of religious texts: A "big data" approach to mind-body concepts in early China. Journal of the american academy of religion, 85(4), 985-1016.

Verma, M. (2017). Lexical analysis of religious texts using text mining and machine learning tools. International Journal of Computer Applications, 168(8), 39-45.

"She Was Simply a Woman Who Was in Search of Variety": Supporting Complex Searches on The Willa Cather Archive with Orchid

Jessica Dussault

jdussault@unl.edu
University of Nebraska-Lincoln, United States of America

Tunink Greg

University of Nebraska-Lincoln, United States of America

Dalziel Karin

University of Nebraska-Lincoln, United States of America

Keywords: search, technical debt, open source, software migration, sustainability
Topics: Global, English, Contemporary, digital research infrastructures development and analysis, software development, systems, analysis and methods, Computer science, Library & information science

"She Was Simply a Woman Who Was In Search of Variety"

Supporting Complex Searches on The Willa Cather Archive with Orchid

Dussault, Jessica; Tunink, Greg; Dalziel, Karin
University of Nebraska-Lincoln

The Center for Digital Research in the Humanities (CDRH)1 is developing the open-source software, Orchid, to provide a browsing, searching, and viewing interface for documents stored in a center-wide API.2 Orchid was built not only to accommodate new projects at the CDRH, but also as a tool to allow rapid upgrading of older sites using TEI-XML3 documents as part of a mission to address the technical debt of dozens of aging sites. In late 2018, the first production site powered by Orchid, The Complete Letters of Willa Cather,4 was launched as a complement to the existing Willa Cather Archive (WCA) website.5 While The Complete Letters was now powered by Orchid, the rest of the WCA remained in Apache Cocoon,6 a framework it had used since the mid-2000s. In 2019, the CDRH challenged itself with migrating the entire WCA to an Orchid-powered application - a strong proof-of-concept that Orchid was equally suited to the task of hosting complex existing projects as well as new sites.

The first obstacle to migrating the WCA site was that Orchid had not been built for the complexity of sites such as the WCA. The site draws from overlapping sets of documents with differing displays and behavior, so the new application required both a site-wide search in addition to the existing Complete Letters search functionality. In order to address this challenge, Orchid logic was altered to support multiple sections of websites with customizable browsing options, search defaults, interfaces, and internationalization.7 This was engineered to re-use as much of the existing Orchid code and require as little additional configuration and code in the application as possible.

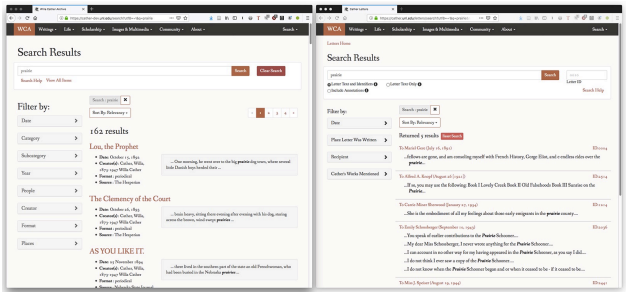


Figure 1: Screenshots of the site-wide WCA search (left) and The Complete Letters search (right). Differences include the ability to search annotations separately, different filters, a custom display, and a search by letter identifier.

This multi-section functionality is predicted to be utilized by complex sites slated for upgrade from older technologies, such as The Walt Whitman Archive.8 Perhaps most exciting of all, the multi-section functionality of Orchid brings us closer to a long-term goal of the CDRH to offer simple interfaces for engaging with older documents from dozens of projects available as sub-sites in a single, centralized application.

Footnotes

1 https://cdrh.unl.edu

2 Source code and documentation available at https://github.com/cdrh/orchid

3 The majority of the CDRH's projects involve text documents marked up in TEI-XML. For more information about this encoding method, see https://tei-c.org/

4 https://cather.unl.edu/letters

5 https://cather.unl.edu

6 https://cocoon.apache.org/

7 These pull requests on the Orchid GitHub repository handled the bulk of the modifications required to create multi-section support: https://github.com/CDRH/orchid/pull/133, https://github.com/CDRH/orchid/pull/143

8 https://whitmanarchive.org/ The Walt Whitman Archive is also powered primarily with Apache Cocoon and needs to be migrated to a newer, actively-supported technology

Sources

Apache Software Foundation (2013). "The Apache Cocoon Project." Apache Software Foundation. https://cocoon.apache.org/

Cather, Willa (1894 October 9). "Amusements." Nebraska State Journal. https://cather.unl.edu/writings/journalism/j00071

Center for Digital Research in the Humanities (2019). "Center for Digital Research in the Humanities." University of Nebraska-Lincoln. https://cdrh.unl.edu/

Center for Digital Research in the Humanities (2019). "Orchid." GitHub. https://github.com/cdrh/orchid

Dalziel, Karin; Dussault, Jessica; Tunink, Greg (2018). "Legacy No Longer: Designing Sustainable Systems for Website Development." DH2018. https://dh2018.adho.org/en/legacy-no-longer-designing-sustainable-systems-for-website-development/

Text Encoding Initiative (2019). "TEI: Text Encoding Initiative", Text Encoding Initiative. https://tei-c.org/

Walt Whitman Archive, The (2019). "The Walt Whitman Archive." Center for Digital Research in the Humanities. https://whitmanarchive.org/

Willa Cather Archive, The (2019). "The Complete Letters of Willa Cather." University of Nebraska-Lincoln. https://cather.unl.edu/letters/

Willa Cather Archive, The (2019). "The Willa Cather Archive." Center for Digital Research in the Humanities. https://cather.unl.edu/

SDC4Lit - a science data center for literature

Claus-Michael Schlesinger
University of Stuttgart

Gabriel Viehhauser
viehhauser@ilw.uni-stuttgart.de
University of Stuttgart

Mona Ulrich
Deutsches Literaturarchiv Marbach

Jan Hess
Deutsches Literaturarchiv Marbach

Roland Kamzelak
Deutsches Literaturarchiv Marbach

Heinz Werner Kramski
Deutsches Literaturarchiv Marbach

Kerstin Jung
University of Stuttgart

Jonas Kuhn
University of Stuttgart

Björn Schembera
University of Stuttgart

Thomas Bönisch
University of Stuttgart

Andreas Kaminski
University of Stuttgart

Keywords: digital literature, archive, analysis, modeling
Topics: Europe, English, Contemporary, digital archiving, text mining and analysis, Literary studies

SDC4Lit - A science data center for literature

Digital media changes the conditions of production, distribution, reception and scientific research of literature. The Digital Humanities currently focus on the possibilities digital methods provide for researching literature (Distant/Scalable Reading). However, digital media has also played an important role in the development of new literary genres and poetics like hypertext novels, literary blogs, twitter bots, text generators and many more. (Hayles 1999; Slater 2018; Rettberg 2019)

On the one hand, these texts seem to be perfectly suited for computational methods since they are genuinely electronic. On the other hand, electronic media implies many new requirements for cultural institutions when dealing with these materials, especially regarding the process of archiving and re-presentation (Seiça, 2015). The high frequency of technical development in digital media means that hardware and software can be outdated and thus hard to find only years after the production of an electronic literary work. This calls for new archival strategies and for new, non-linear models of electronic text. (Kramski/ Buelow, 2011; Connor et al., 2019)

The newly founded interdisciplinary Science Data Center for Literature (SDC4Lit) aims to systematically reflect the technical and aesthetic nature and the resulting requirements of digital literature. Research and development take into account that modeling, archiving and analyzing digital literature needs to be closely connected to research questions in the respective domains. The SDC4Lit project combines four partners: the German Literature Archive (DLA), the High-Performance Computing Center Stuttgart, the department of computational linguistics and the department of digital humanities, both University of Stuttgart.

SDC4Lit focuses on archival solutions and research methods. The group will develop longterm repositories for digital literature and respective research data and a web platform providing access to the repositories and to computational research tools. Thus, the project will develop and implement a sustainable data life cycle for born digital literature within its institutional framework. The repositories represent the main storage for continuous harvesting of net literature and for other born-digital material. The research platform provides specific tools and methods for further analysis of the works in the repositories. Repositories and research platform will be connected through an Advanced Programming Interface (API). This approach provides the possibility to integrate more repositories and to provide low-level access to the SDC-repositories to other projects. In addition to the digital objects and their metadata the project will develop and implement a dedicated repository for all research data. This includes all research data produced by the project SDC4Lit itself as well as research data generated by users, e.g. annotations or additional metadata for single objects or object classes.

The involvement of the communities participating in the production, distribution, research and mediation of electronic literature will be a decisive element for the project. Outreach measures include workshops and seminars, working with focus groups and the installation of an advisory board. The collection, provision, research and mediation of literature in electronic media is a task that concerns both research and archives. It is the goal of SDC4Lit to work on and fulfill these tasks in an inclusive and interdisciplinary way as described above.

Bibliography

Connor, Michael, Aria Dan, Dragan Espenschied (Ed.) (2019): The Art Happens Here. Net Art Anthology, New York: New Museum of Contemporary Art.Hayles, N. Katherine (1999): How we became posthuman : virtual bodies in cybernetics, literature, and informatics, Chicago (Ill.): The University of Chicago Press.Kramski, Heinz Werner, Ulrich von Bülow: "Es füllt sich der Speicher mit köstlicher Habe" - Erfahrungen mit digitalen Archivmaterialien im Deutschen Literaturarchiv Marbach ", in: Caroline Y. Robertson von Trotha. Robert Hauser (Ed.), Neues Erbe : Aspekte, Perspektiven und Konsequenzen der digitalen Überlieferung, Karlsruhe: KIT Scientific Publishing 2011, p. 141-162.Rettberg, Scott (2019): Electronic literature. Cambridge, UK: Polity Press.Seiça, Álvaro (2015): Um Feixe Luminoso: Uma Leitura da Coleção de Literatura Electrônica Portuguesa. Florianópolis: Universidade Federal de Santa Catarina.Slater, Avery (2018): "Crypto-Monolingualism. Machine Translation and the Poetics ofAutomation". In: Amodern. 8.

An Improvement of Collaborative Digital Scholarly Edition with IIIF

Kiyonori Nagasaki
nagasaki@dhii.jp
International Institute for Digital Humanities, Japan

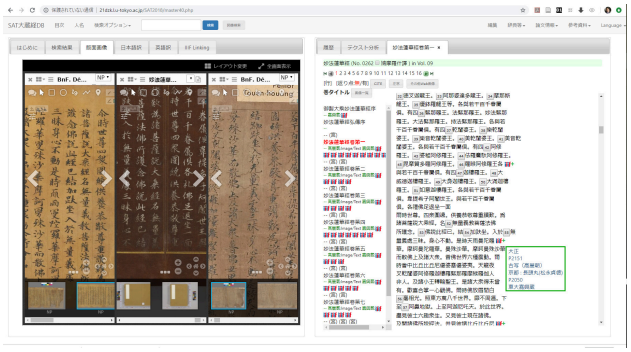
Ikki Ohmukai
The University of Tokyo

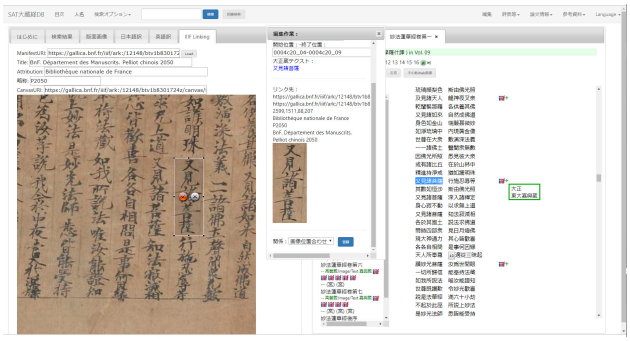
Toru Tomabechi
International Institute for Digital Humanities, Japan

Masahiro Shimoda
International Institute for Digital Humanities, Japan

Keywords: Buddhist Studies, TEI, Critical apparatus, digital facsimiles
Topics: Asia, English, 5th-14th Century, 15th-17th Century, 18th Century, digital research infrastructures development and analysis, scholarly editing and editions development, analysis, and methods, Asian studies, Philology

Digital scholarly edition (henceforth, DSE) has been gradually improved as structured electronic texts through TEI (Text Encoding Initiative) Guidelines[1] since around three decades ago in order to find better solution for not only digital humanities but also humanities in general. Emergence of IIIF (International Image Interoperability Framework)[2] has recently been providing new possibilities to it via its interoperable nature. This presentation will report on an improvement of our collaborative digital scholarly edition as a use case of utilizing IIIF which enables us to manipulate hi-resolution images provided from many cultural institutions in the world such as Gallica, Cambridge University Library, National Diet Library of Japan, and Kyoto University for DSE.





Revue 2.0 Repenser: la mission des revues savantes en sciences humaines et sociales

Antoine Fauchié
antoine.fauchie@umontreal.ca
Chaire de recherche du Canada en écritures numériques, Canada

Margot Mellet
Chaire de recherche du Canada en écritures numériques, Canada

Servanne Monjour
Chaire de recherche du Canada en écritures numériques, Canada

Nicolas Sauret
Chaire de recherche du Canada en écritures numériques, Canada

Marcello Vitali-Rosati
Chaire de recherche du Canada en écritures numériques, Canada

Keywords: academic publishing, journals, open access, digital textualities, publishing protocols
Topics: Europe, French, North America, Contemporary, digital publishing projects, systems, and methods, scholarly editing and editions development, analysis, and methods, Book and print history, Media studies

Quels savoirs pour demain ? À l'issue de la Seconde Guerre mondiale, on assiste à un progressif phénomène de commercialisation du savoir qui change la nature de ce dernier (Guédon et Loute 2017). L'intervention du numérique ouvre à nouveau des possibilités de changement du savoir, tout en en accentuant l'aspect commercial (Kaplan 2014). La présence des grands acteurs commerciaux du web, que l'on peut cerner approximativement avec l'acronyme GAFAM (Google, Apple, Facebook, Amazon, Microsoft), ainsi que la montée en puissance de cinq éditeurs académiques en SHS (Elsevier, Taylor&Francis, Wiley-Blackwell, Springer, Sage, cf. Larivière, Haustein, et Mongeon 2015), sont une manifestation claire de cette intensification opérée par la conjonction du numérique et du commercial : elle produit un effet quasi-hégémonique sur la connaissance et les savoirs en les assujettissant à une uniformisation linguistique et épistémologique sans précédent (Sordi et Fiormonte 2019). Dans ce cadre, l'hégémonie de l'anglais et d'une vision du monde propre à certaines élites nord-américaines (Morozov 2012, 2013) est déjà la réalité de notre époque et menace de s'amplifier dans un futur proche (Wormser 2018).Comment contrer ce modèle ? Comment garantir que les savoirs de demain soient fondés sur une pluralité de visions du monde, de langues et de cultures ? Comment faire en sorte que ces savoirs ne soient pas orientés par les intérêts commerciaux, et conservent bien une dimension publique et collective ?Ces questions sont au centre du projet de développement partenariat Revue 2.0 (dont les résultats préliminaires sont disponibles sur le site http://revue20.org). Cette vaste enquête, menée en partenariat avec les principaux acteurs de l'édition savante francophone en sciences humaines (SH), nous a démontré que les revues, en tant que reflet de communautés de recherche, constituaient de véritables lieux d'expérimentation de nouvelles formes de production du savoir. Historiquement, elles ont su s'adapter à différents paradigmes. Elles ont été des ateliers de la pensée dans le monde de l'imprimé en créant par exemple des formats de diffusion originaux selon le contexte matériel ; dans un domaine comme celui des SH où la forme monographique reste le support privilégié et le plus institutionnel pour diffuser les résultats d'une recherche (Larivière et Lobet 2018), les revues ont su expérimenter et innover, ouvrir des espaces autres, permettre la conversation et le dialogue au sein des communautés scientifiques (Vitali-Rosati 2018). Elles ont garanti l'existence d'une multiplicité d'approches, de modèles épistémologiques et de visions du monde. Or, là où l'on pouvait penser que les revues devaient mieux opérer leur remédiation pour s'adapter à l'environnement numérique, nous avons au contraire compris grâce à l'enquête Revue 2.0, que de par leur multiplicité et leur diversité, les revues peuvent contribuer à la richesse et à la diversité des environnements numériques, aujourd'hui constitués en un espace hégémonique.Le projet Revue 2.0 vise à repenser le rôle des revues savantes dans les sciences humaines et sociales et à accompagner les acteurs de l'édition scientifique dans leur transition numérique, pour faire en sorte que les revues savantes en SH restent les protagonistes des savoirs de demain, en garantissant ainsi dans l'espace numérique une pluralité de modèles, de langues et de cultures, tout en préservant la dimension publique des savoirs.Les revues sont un terrain d'expérimentation privilégié pour imaginer de nouveaux modes de production, diffusion et légitimation du savoir à l'époque du numérique (Vitali-Rosati 2014) qui permettent de relever un défi décisif : préserver la possibilité d'une connaissance multiple et plurielle et rester les acteurs au fondement des différents modèles épistémologiques qui structurent le savoir.Au début des années 1990, la naissance des premiers diffuseurs scientifiques numériques comme Muse a été suivie dans le domaine francophone par la création d'OpenEdition ou d'Érudit. Il était alors question de convaincre les revues d'investir les environnements numériques. Vingt ans plus tard, comme nous l'a démontré le projet Revue 2.0, ce passage a été accompli, et les revues se sont inscrites dans des écosystèmes de diffusion qui, par contre, ne prennent pas suffisamment en compte leurs spécificités. Les revues réalisent aujourd'hui que pour conserver leur diversité éditoriale et épistémologique, il est nécessaire de questionner ces écosystèmes et d'imaginer de nouveaux modèles, en évitant l'homogénéisation. Si les vingt dernières années ont été celles du passage au numérique, les dix prochaines seront celles de l'émergence de nouveaux régimes de connaissance et de la diversification des espaces numériques.

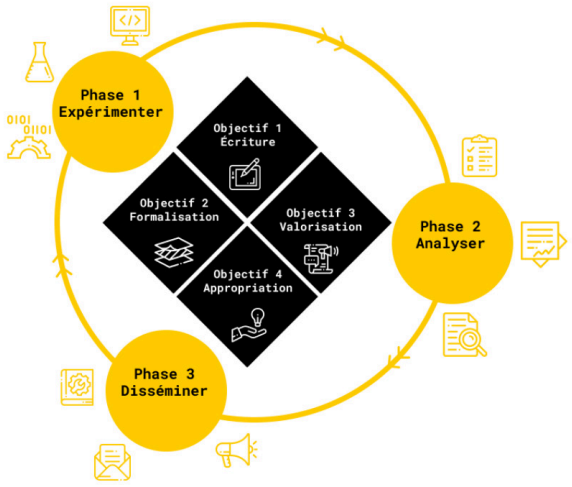


Figure 1: Schéma global du projet

Bibliographie

Guédon, Jean-Claude, et Alain Loute. 2017. « L'histoire de la forme revue au prisme de l'histoire de la grande conversation scientifique . Entretien avec Jean-Claude Guédon réalisé par Alain Loute, préparé avec l'aide de Caroline Glorie, Thomas Franck et Andrea Cavazzini. » Cahiers du GRM. publiés par le Groupe de Recherches Matérialistes Association, n? 12 (décembre). https://doi.org/10.4000/grm.912.Kaplan, Frédéric. 2014. « La question de la langue à l'époque de Google ». In Digital Studies : organologie des savoirs et technologies de la connaissance, 189. Innovation. FYP éditions.Larivière, Vincent, Stefanie Haustein, et Philippe Mongeon. 2015. « The Oligopoly of Academic Publishers in the Digital Era ». PLOS ONE 10 (6):e0127502. https://doi.org/10.1371/journal.pone.0127502.Larivière, Vincent, et Delphine Lobet. 2018. « La mort des livres dans les sciences humaines et sociales, et en arts et lettres? » Acfas. Acfas. 14 juin 2018. https://www.acfas.ca/publications/decouvrir/2018/06/mort-livres-sciences-humaines-sociales-arts-lettres.Morozov, Evgeny. 2012. The Net Delusion: The Dark Side of Internet Freedom. Reprint edition. PublicAffairs.Morozov, Evgeny. 2013. To Save Everything, Click Here: The Folly of Technological Solutionism. First Trade Paper Edition edition. PublicAffairs.Sordi, Paolo, et Domenico Fiormonte. 2019. « Geopolitica Della Conoscenza Digitale. Dal Web Aperto All'impero Di GAFAM. » DigitCult - Scientific Journal on Digital Cultures 4 (1):21?36. https://doi.org/10.4399/97888255263183.Vitali-Rosati, Marcello. 2014. « Les Revues Littéraires En Ligne : Entre Éditorialisation et Réseaux d'intelligences ». Études françaises 50 (3):83. https://doi.org/10.7202/1027191ar.Vitali-Rosati, Marcello. 2018. « À Quoi Servent Les Publications Scientifiques? » La vie de la recherche scientifique, n? 412:19?22. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/21016.Wormser, Gérard. 2018. Facebook. Les atelier de Sens public.

How Issues Guide The Life of News Stories

Blanca Calvo Figueras
b.calvo.figueras@student.rug.nl
University of Groningen

Tommaso Caselli
University of Groningen

Marcel Broersma
University of Groningen

Keywords: agenda setting, issue type, clustering, plot structure
Topics: Global, Europe, English, Contemporary, data modeling, text mining and analysis, Cultural studies, Media studies

How Issues Guide The Life of News Stories

Blanca Calvo Figueras, Tommaso Caselli, Marcel Broersma
University of Groningen
b.calvo.figueras@student.rug.nl
t.caselli | m.j.broersma@rug.nl

Abstract

Explaining the dynamics between the issues mass media emphasize and the salience audiences attribute to them is the main concern of agenda-setting research. Previous studies (Scheufele and Tewksbury 2006; Yunjuan et al. 2019; Wanta and Ghanem 2006) suggest that the capacity of media to influence the salience of issues depends on four parameters, namely: (1) obtrusiveness; (2) duration; (3) abstractness; and (4) dramatism. Soroka (2002) built upon these parameters by developing an issue typology:

- Prominent issues affect a relevant number of people directly (e.g. the rise of salaries) and are thought to leave little room for media impact on public opinion.
- Sensational issues are initiated by a dramatic event (e.g. a kidnapping) but have little observable

impact on the everyday life of the majority of the population.

- Governmental issues are perceived as abstract or as not having direct effect on people's lives (e.g. public debt) and do not offer dramatic or exciting components.

Assuming the validity of Soroka's typology, we ask whether and how issue-types give rise to different life patterns of news stories.

We thus identify three measurable dimensions: (a.) lifespan (i.e., the number of days over which articles of a news story are published); (b.) intensity (i.e., the number of articles per day per story); and (c.) burstiness (i.e., the speed with which stories go from emerging to their climax).

We collected a corpus of 50,385 political articles from major Spanish newspapers in 2018. News stories were generated by aggregating articles with K-means clustering. For each week, we identified the number of clusters (i.e. stories) using the elbow method (k ranging between 1 and 30), and removed general clusters using silhouette analysis.

A subset of the clusters was evaluated against manually labeled data by checking the extent to which clusters contain a single class (i.e., purity), obtaining a score of 0.87. Afterwards, we assigned each story to one issue-type.

The quantitative dimensions show that the issue-types differentiate only to some extent. Sensational stories have the highest intensity, shortest lifespan, and highest burstiness (i.e., they appear "out of the blue"). Prominent stories last for long periods (14 days on average) but have low intensity and burstiness, i.e. they slowly grow to their climax. Governmental stories lay in the middle between these two latter: they are similar to Sensational stories in intensity, and to Prominent stories in lifespan. However, their burstiness score suggests variations in the speed of growth.

Figure 1 visualizes the patterns of the stories aggregated by type[1]: sensational stories have one climax event and long falling actions, while Prominent stories may have multiple climax events. The different typologies seem to instantiate different plot structures (Bal 1997).

The results of our empirical analysis indicate that different issue-types exhibit different behaviors when represented as news stories. To the best of our knowledge, this is the first attempt to operationalize Soroka's framework in a digital approach. Future research will pursue understanding the capacity of media agents to define public agenda through timing.

Our approach is language independent and can be applied to study news issue-types across media platforms and countries[2].

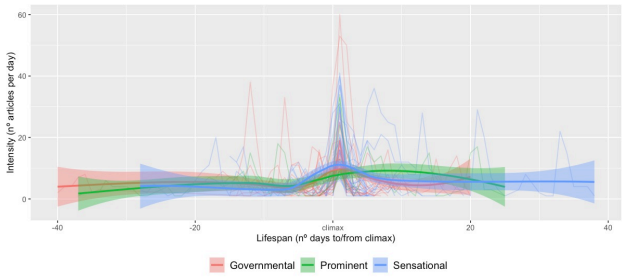


Figure 1: All the stories of our corpus are plotted with its climax in 0 (transparent lines). The smoothed conditional mean of each type is plotted in bold.

References

Bal, M. 1997. Narratology: Introduction to the Theory of Narrative. University of Toronto Press.

Scheufele, Dietram, and David Tewksbury. 2006. "Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models." Journal of Communication 57 (1): 9-20.

Soroka, Stuart Neil. 2002. Agenda-Setting Dynamics in Canada. UBC Press.

Wanta, Wayne, and Salma Ghanem. 2006. "Effects of Agenda Setting." In Mass Media Effects Research: Advances Through Meta-Analysis. Routledge.

Yunjuan, Luo, Hansel Burley, Alexander Moe, and Mingxiao Sui. 2019. "A Meta-Analysis of News Media's Public Agenda-Setting Effects, 1972-2015." Journalism & Mass Communication Quarterly 96 (1): 150-72.

[1] An interactive application with all the individual stories is available: https://bit.ly/2Pbg438

[2] Our methodology is open source and can be found here: https://bit.ly/2X95fTt

Communicating the TEI Across Linguistic and Cultural Boundaries

Hugh Cayless
philomousos@gmail.com
DC3, Duke University

Martina Scholger
Centre for Information Modeling - Austrian Centre for Digital Humanities, University of Graz

Helena Bermúdez Sabel
Université de Lausanne

Luis Meneses
Electronic Textual Cultures Laboratory, University of Victoria

Gimena del Rio Riande
Consejo Nacional de Investigaciones Científicas

Kiyonori Nagasaki
International Institute for Digital Humanities

Keywords: translation, multilingualism, community-building, Text Encoding Initiative, internationalization
Topics: Global, English, Contemporary, sustainable procedures, systems, and methods, text encoding and markup language creation, deployment, and analysis, Humanities computing

The Text Encoding Initiative [1] is one of the oldest continuously developed DH projects in existence. For many years, efforts have been made to translate the TEI Guidelines and specifications into languages other than the original English [2][3][4]. These efforts, however, have often been frustrated or eventually abandoned. Despite the TEI's ingenious infrastructural support for translation, for which much credit is due to the late Sebastian Rahtz, it is difficult for translation efforts to stay continuously up to date with the Guidelines, which change incrementally and employ terminology that is often technical and sometimes unclear. In a recent meeting as part of the Triangle Scholarly Communications Institute, a group of scholars focused on the problem of communicating the TEI and developed proposals to remedy the problems with translating the TEI documentation [5]. The conclusions and next steps that were identified in that meeting are the focus of this poster.

It is quite clear that the flow of knowledge as we try to communicate the TEI cannot simply go from English to other languages: the act of translating can highlight cases where the Guidelines and specifications are insufficiently clear or even wrong. Revisions to the original document may be required. Research groups working with different languages may discover they have problems and solutions in common. Therefore, it is important that the TEI Consortium should take active responsibility for coordinating ongoing translation efforts, foregrounding them, and ensuring that they have the tools and affordances they need for their work. Additionally, the advent of reasonably useful machine translation may prove to be an important catalyst (though not by itself sufficient) for translation efforts. A glossary of TEI technical terms must be developed, accompanied by notes and examples for translators, s o
that it is clear not just what a term means, but why it (o r its translation) was chosen, and how it is to be used.

Further, the affordances on the TEI's website must be improved so that it is clear to beginners where to start and where useful examples may be found. As part of this, and of the translation of the specifications, it may be necessary to move the Guidelines themselves out of the foreground and to make the more easily internationalized specifications the primary point of contact for users. Although it is probably inevitable that a single language (which must probably be English for historical reasons) be used as a common basis of discourse for the TEI, a concerted effort must be made to provide introductory materials in other languages [6][2]. This poster will document the steps we are taking towards ensuring that communication between the different linguistic communities using TEI can flow in multiple directions. We intend to make the ways TEI is communicated more equitable so that it may be adopted by a more diverse user base.

References

[1] TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. 3.6.0. July 2019. TEI Consortium. http://www.tei-c.org/Guidelines/P5 (2019-10-16).

[2] Allés Torrent, Susanna / Calarco, Gabriel / Rio Riande, Gimena del. TTHub. Text Technologies Hub. Recursos sobre tecnologías del texto y edición digital: https://tthub.io/ (2019-10-16).

[3] Stadler, Peter / Scholger, Martina (2016): Report on the TEI2German Translatathon at Vienna 2016. https://tei-c.org/tools/i18n/tei2german (2019-10-17).

[4] SIG East Asian/Japanese, TEI Consortium: https://github.com/TEI-EAJ (2019-10-16).

[5] Bermúdez Sabel, Helena / Cayless, Hugh / Meneses, Luis / Nagasaki, Kiyonori / Rio Riande, Gimena del / Scholger, Martina (2019): https://trianglesci.org/2019/07/17/communicating-the-tei-to-a-multilingual-user-community (2019-10-16).

[6] Terras, Melissa / Van den Branden, Ron / Vanhoutte, Edward (2017): TEI by Example. http://www.teibyexample.org (2019-10-16).

The Otletosphere: an interactive map of the network around Paul Otlet

Olivier Le Deuff
oledeuff@gmail.com

Université Bordeaux Montaigne, laboratoire MICA E3D

Arthur Perret
arthur.perret@u-bordeaux-montaigne.fr
Université Bordeaux Montaigne, laboratoire MICA E3D

Clement Borel
Université Bordeaux Montaigne, laboratoire MICA E3D

Keywords: Paul Otlet, information science, network, visualisation
Topics: Global, Europe, English, 19th Century, 20th Century, digital biography, personography, and prosopography, Interface design, development, and analysis, Library & information science

The Otletosphere is a relational mapping of personalities and institutions related to Paul Otlet. This project is based on Paul Otlet's strong involvement in international organisations for peace as well as in bibliographic and documentary institutions.

The Otletsphere runs in a web interface (Figure 1) and is built in HTML, CSS and Javascript. It makes use of the Vis.js and Tabletop.js libraries for data handling and visualization. Data is stored in JSON.

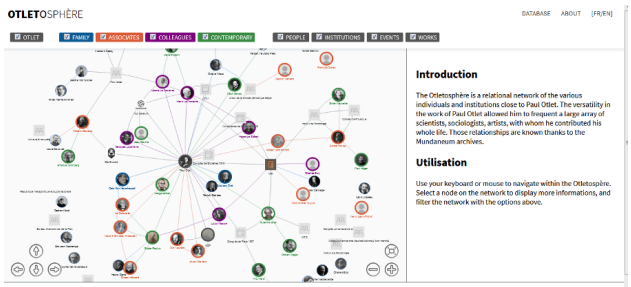


Figure 1

Because the world is multidimensional: Annotations for storytelling with 3D objects

Zoe Schubert
zoe.schubert@uni-koeln.de
Institut für Digital Humanities, Universität zu Köln

Enes Türko#lu
Institut für Digital Humanities, Universität zu Köln
Institut für Medienkultur und Theater, Universität zu Köln

Øyvind Eide
Institut für Digital Humanities, Universität zu Köln

Jan Gerrit Wieners
Institut für Digital Humanities, Universität zu Köln

Kai Michael Niebes
Institut für Digital Humanities, Universität zu Köln

Keywords: annotation, storytelling, 3D modelling
Topics: Global, Europe, English, 20th Century, Contemporary, annotation structures, systems, and methods, digitization (2D & 3D), Humanities computing, Performance Studies: Dance, Theatre

Annotation is a well established method in digital humanities. For textual documents, inline (typically TEI-XML) and stand-off (e.g. character index table based) markup are two widely used approaches. For adding textual annotations, various software affording different types and levels of abstraction are used, including XML editors such as Oxygen [1] as well as online systems such as Catma [2]. Many annotation systems enable annotation of other media than verbal text in various ways, often with a focus on images and figures. Systems are also available for audio annotation of text and music, and for video annotations [3].

Annotation systems for 3D models are less common and many users in the humanities and cultural heritage resort to online platforms such as Sketchfab [4]. Commercial platforms might be a good choice in many cases, but copyright issues and legal obligations make them unwanted or unusable in other. Furthermore, with these 3D presentation platforms [5], the focus is not on (collaborative) annotation as a method. Since 2017, the Departments for Digital Humanities and Media Culture and Theatre at the University of Cologne, in cooperation with the university's Theatre Collection, has developed Kompakkt [6], a free online tool [7] for multi-modal 3D annotations which is specifically designed for collaborative work and storytelling. Kompakkt is fully open source [8] and opt for new ways to provide public access for cultural heritage objects, archival material, etc.

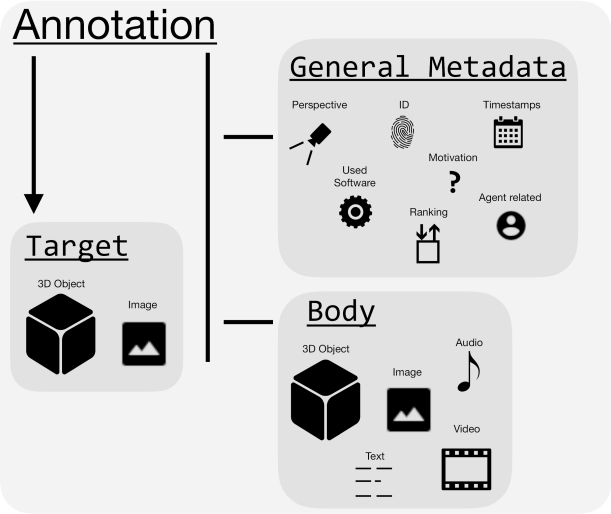


Figure 1: A schematic representation of annotations in Kompakkt [9]

With Kompakkt, the 3D representation of an object serves as the hub of an open-ended collection of heterogeneous information established through the use of multimodal annotations, cf. Figure 1. Through personalised and group level collections of digital objects, Kompakkt enables a novel solution for gathering, generating and inter-connecting information about digital objects and related non-digital things. Registered users may decide if their content is restricted, private or public to control the visibility and access to their objects.

Through opening up for collaborative annotation of objects, multi-vocal voices can be heard, presenting traditional scholarly and cultural heritage management perspectives as well as knowledge based on citizen humanities. Diverse groups around the world can annotate, discuss, and tell multiple stories about collections of objects, making sets of open data available to be discussed, further developed, and experienced [10].

The annotation functionality for images and 3D models enables linking between repository-internal objects and external information through weblinks. Therefore every single point on the surface of the mesh can be selected. An annotation is not only linked to a specific location in space, it also includes and covers a user defined related perspective, the corresponding point of view. Multiple Annotations in Kompakkt are chained in flexible series what allows users to walk through the different related, thereby enabling a camera track by stepping from one annotation to the next and its corresponding perspective. Consequently, presentations and storytelling can be implemented through a series of annotations traversing through different objects.

The challenge posed by the future of Kompakkt is primarily to increase the community and scale up the number of available objects. Therefore it is not only important to ensure the quality of the data through curation but also to develop the import and export functionality further and in general to enhance the usability with the help of user feedback. Additionally with the feedback from the community from the prototyping phase helps to improve and expand the annotation functionality and implement new suggested features.

References

[1] Oxygen XML Editor - The Complete Solution for XML Authoring, Development & Collaboration, .

[2] CATMA - For Undogmatic Textual Markup And Analysis, .

[3] For instance: Praat: Doing Phonetics By Computer, .

[4] Sketchfab - Publish & Find 3D Models Online, .

[5] See also: 3DHOP, .

[6] An online version for testing Kompakkt is available from: .

[7] Based on Web Technologies: Babylon.Js, , and Angular, .

[8] The source code is available on Github; DH-Cologne/Kompakkt,.

[9] The underlying data model is based on the W3C Recommendation Web Annotation Data Model, .

[10] To that end, Kompakktnot only allows the export of the according metadata and annotations of a given object, but also is able to pull and use data from Europeana Collections to encourage interoperability.

Critical Making and Public Facing DH in the Liberal Arts Classroom a case study.

Amanda Henrichs
ahenrichs@amherst.edu
Amherst College, United States of America

Keywords: liberal arts, college, pedagogy, public
Topics: EnglishNorth America20th Century, Contemporary, curricular and pedagogical development and analysis, public humanities

collaborations and methods, Education/ pedagogyMedia studies

Digital humanities have historically been underrepresented at small liberal arts colleges. Though this is changing, the pace is slow, and most DH pedagogy remains targeted at graduate students and faculty at R1 universities. Bryan Alexander and Rebecca Frost Davis have identified several reasons for the lack of DH pedagogy at small liberal arts colleges (SLACs), including "logistics, infrastructure, and campus identity."[1] Briefly, SLACs frequently do not have either the critical mass or the funding to make DH courses (which can be resource-intensive for both individual faculty and the broader institution) viable. Additionally, SLACs emphasize a liberal undergraduate education, which can be skeptical of or even hostile to training students in the concrete technical skills that digital humanists frequently need to build digital projects. Nearly a decade after Alexander and Davis identified the particular challenges for DH instruction in a SLAC context, SLACs are shifting towards developing DH minors, certificates, and sometimes even majors in their curricula. Yet many of the challenges-of logistics and campus identity in particular-remain the same. In other words, small colleges can feel insular and insulated from larger digital humanities contexts.

This poster presents a case study of what one response to these challenges can be: to teach students not only how to apply skills like HTML in a particular course context, but also to frame DH and DH pedagogy as always aimed toward a wider audience. The course which serves as the case study for the poster, to be taught in Spring 2020 at Amherst College, is titled "Books and Their Afterlives: Text as Technology." Students both examine and use a wide variety of textual technologies from across the centuries: they handle rare books in a college archive, write HTML to develop a narrative in Twine, and experiment with 3D printing pens and paper circuits in order to investigate text as a concept, a technology, and a material artefact.

This poster serves two functions: first, it will enact practical DH pedagogy by engaging students in developing the poster, encouraging them to think of their course work as public facing. Second, the poster will demonstrate the arc of the course to that public audience. It will include the guiding principles of the course's design: the remaining poster content will be determined by student collaborators.

This poster will be of interest to faculty and instructional staff in environments that might resist DH pedagogy, such as small liberal arts colleges and libraries. It will also be of interest to those in environments that have embraced DH pedagogy who might want an example of engaging students in active learning through public-facing digital work.

[1] "'Chapter 21: Should Liberal Arts Campuses Do Digital Humanities? Process and Products in the Small College World | Bryan Alexander and Rebecca Frost Davis' in 'Debates in the Digital Humanities' on Manifold," Debates in the Digital Humanities, accessed October 21, 2019, https://dhdebates.gc.cuny.edu/read/untitled-88c11800-9446-469b-a3be-3fdb36bfbd1e/section/d61cb020-c749-405a-ab59-63f8da332425#ch21.

TextAnnotator: A web-based annotation suite for texts

Giuseppe Abrami
s3676959@stud.uni-frankfurt.de
Goethe University Frankfurt, Germany

Alexander Mehler
Goethe University Frankfurt, Germany

Stoeckel Manuel
Goethe University Frankfurt, Germany

Keywords: Annotation, Inter-Annotator-Agreement, UIMA, NLP
Topics: Global, English, Contemporary, annotation structures, systems, and methods, software development, systems, analysis and methods, Computer science, Humanities computing

The annotation of natural language texts and their use is addressed in many projects in the digital humanities. This not only involves the generation of training data, but also the correction of errors by automatic preprocessing. Nowadays there are many methods for automatic text analysis, and just as many tools (e.g. [10, 13, 3, 9]) which encapsulate them for different natural languages as well as for different programming languages. However, there are relatively few annotation tools (e.g. [6, 13, 4, 11]) for correcting annotations or generating training data. The annotation tools mentioned usually only allow a simple annotation of texts as well as a simple visual annotation support. In addition, the use of

knowledge databases, such as Wikipedia, Wikidata, Geonames or similar, are rarely usable. Furthermore, the administration of Corpora, the use of different annotation views, the simultaneous and collaborative annotation of the same texts by different users, the user and group-related granting of access permissions to texts, as well as the dynamic determination of Inter-Annotator-Aggreements, are almost non-existent. However, this limited use of annotation tools shows a gap in the large field of digital humanities that can be closed by the so-called TextAnnotator [8].

The TextAnnotator includes a variety of modules for the annotation of texts, which contains the annotation of argumentative, rhetorical, propositional and temporal structures as well as a module for named entity linking and rapid annotation of named entities (Fig. 2). Especially the modules for annotation of temporal, argumentative and propositional structures are currently unique in web-based annotation tools [2].

TextAnnotator, which allows the annotation of texts as a platform, is divided into a front- and a backend component. The backend is a web service based on WebSockets, which integrates the UIMA Database Interface [1] to manage and use texts. UIMA [5] acts as de facto standard for all NLP tasks and almost all preprocessing tools produce a UIMA output. In order to use raw texts and preprocessed texts with TextAnnotator, they first are automatically converted into the UIMA format with the help of the so-called TextImager [9] and preprocessed.

In addition, texts are made accessible by using the ResourceManager and the AuthorityManager, based on user and group access permissions [7] (Fig. 1). The use of different components allows the flexible and project-related use of one tool for different purposes. Therefore, texts can be placed in a flexible folder structure and edited by different teams. In addition, different views of a document can be created and used depending on the scenario.

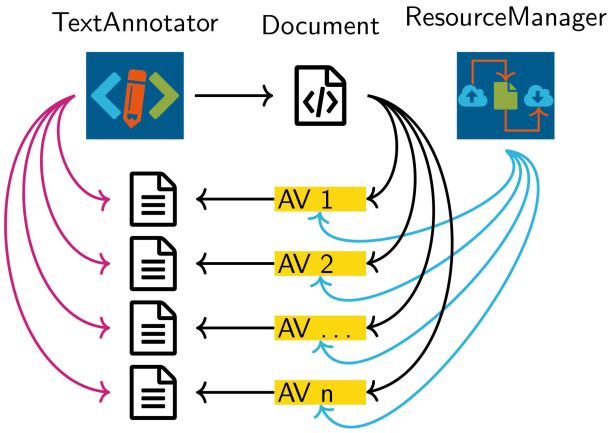


Figure 1: Schematic diagram of the use of annotation views (AV). TextAnnotator has access on documents that contain annotation views which are accessible to users. By this assignment TextAnnotator uses the annotations in the individual views for annotation through implemented tools.

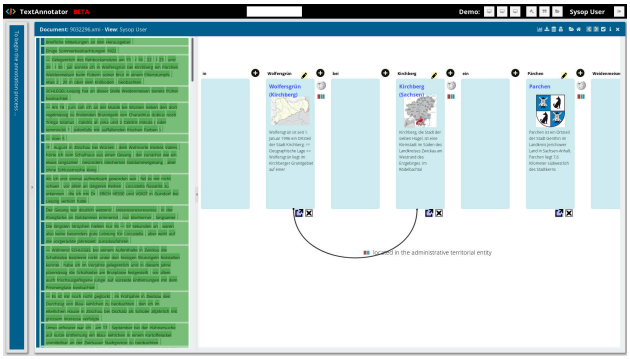


Figure 2: Extract from an annotation session and the use of KnowledgeBaseLinker. The individual tokens can be linked to knowledge resources or the entries can be modified. In this scenario the texts were already automatically preprocessed by the TextImager and the lower line shows an implicit relation which was interpreted based on the Wikidata entries of the respective assignments to the knowledge database Wikipedia.

Through the use of the frontend component, developed in ExtJS, browser-based access to the texts and the available annotation tools is enabled. Once a document has been opened, access is gained to the annotations stored within annotation views in which these are organized. (Fig. 3). Any annotation view can be assigned with access permissions and by default, each user obtains his or her own user view for every annotated document. In addition, with sufficient access permissions, all annotation views can also be used and curated (Fig. 3). This allows the possibility to calculate

an Inter-Annotator-Agreement [12] for a document, which shows an agreement between the annotators. Annotators without sufficient rights cannot display this value so that the annotators do not influence each other. This contribution is intended to reflect the current state of development of TextAnnotator, demonstrate the possibilities of an instantaneous Inter-Annotator-Agreement and trigger a discussion about further functions for the community.

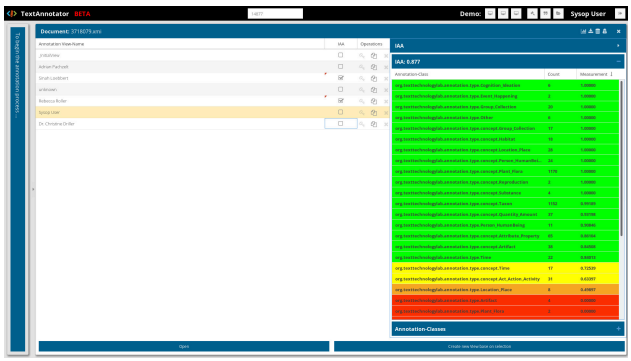


Figure 3: A open document is shown in TextAnnotator. The annotation views are displayed on the left. User views are named with the names of the users other views are annotation views that can be authorized by the user. The IAA value (right) of the document is visualized based on the selected annotation views (left) and the previously selected annotation classes. The agreement of the annotations can be modified arbitrarily (views, classes) and is calculated directly. At the same time, the agreement is highlighted in different colors.

[1] Abrami, Giuseppe and Alexander Mehler (2018). "A UIMA Database Interface for Managing NLP-related Text Annotations". In: Proc. of LREC. Miyazaki, Japan.

[2] Abrami, Giuseppe, Alexander Mehler, Andy Lücking, Elias Rieb, and Philipp Helfrich (2019). "TextAnnotator: A flexible framework for semantic annotations". In: Proc. of ISA-15. Gothenburg, Sweden.
[3] Castilho, Richard Eckart de and Iryna Gurevych (2014). "A broad-coverage collection of portable NLP components for building shareable analysis pipelines". In: Proc. of OIAF4HLT. Dublin, Ireland, pp. 1–11

[4] Castilho, Richard Eckart de, Éva Mújdricza-Maydt, Seid Muhie Yimam, Silvana Hartmann, Iryna Gurevych, Anette Frank, and Chris Biemann (2016). "A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures". In: Proc. of COLING, Osaka, Japan. pp. 76–84.

[5] Ferrucci, David, Adam Lally, Karin Verspoor, and Eric Nyberg (2009). Unstructured Information Management Architecture (UIMA) Version 1.0. OASIS Standard.

[6] Gerdes, Kim (2013). "Collaborative Dependency Annotation". In: Proc. of DepLing. Prague, Czech Republic, pp. 88–97

[7] Gleim, Rüdiger, Alexander Mehler, and Alexandra Ernst (2012). "SOA implementation of the eHumanities Desktop". In: Proc. of Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities, Hamburg, Germany.

[8] Helfrich, Philipp, Elias Rieb, Giuseppe Abrami, Andy Lücking, and Alexander Mehler (2018). "TreeAnnotator: Versatile Visual Annotation of Hierarchical Text Relations". In: Proc. LREC. Miyazaki, Japan.

[9] Hemati, Wahed, Tolga Uslu, and Alexander Mehler (2016). "TextImager: a Distributed UIMA-based System for NLP". In: Proc. of the COLING System Demonstrations. FedCSIS. Osaka, Japan.

[10] Hinrichs, Erhard, Marie Hinrichs, and Thomas Zastrow (2010). "WebLicht: Web-based LRT Services for German". In: Proc. of ACL 2010 System Demonstrations. ACLDemos. Uppsala, Sweden, pp. 25–29

[11] Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych (2018). "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation". In: Proc. COLING System Demonstrations. ACL, pp. 5–9

[12] Meyer, Christian M., Margot Mieskes, Christian Stab, and Iryna Gurevych (2014). "DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement". In: Proc. of COLING System Demonstrations. Dublin, Ireland, pp. 105–109.

[13] Stenetorp, Pontus, Sampo Pyysalo, Goran Topi?, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii (2012). "BRAT: A Web-based Tool for NLP-assisted Text Annotation". In: Proc. of EACL. Avignon, France, pp. 102–107.

The Historic Graves crowdsourcing project: 10 years transcribing Irish history

Maurizio Toscano
maurizio.toscano@gmail.com
Universidad de Granada

John Tierney
Eachtra Archaeological Projects

Keywords: Crowdsourcing, Ireland, Cultural Heritage, Public History, Collaborative platforms
Topics: Global, Europe, English, 18th Century, 19th Century, 20th Century, crowdsourcing, public humanities collaborations and methods, History

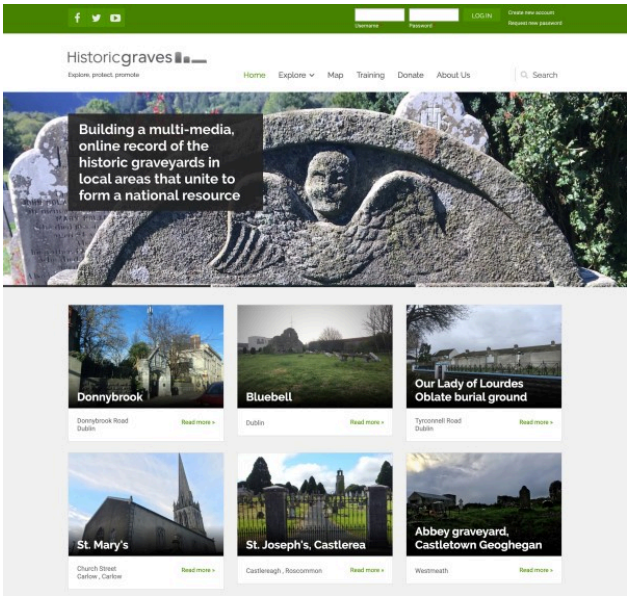
Irish diaspora is how we refer to the historical process of migration from Ireland, recorded since the Early Middle Ages, but particularly evident since the XVIII century. By the 21st century, according to figures in the Irish Emigration Patterns and Citizens Abroad report published in 2017 by the Irish Department of Foreign Affairs and Trade, an estimated 70 million people worldwide claimed some Irish descent, that largely maintain a connection with Irish cultural identity and heritage.

The online project Historic Graves (https://historicgraves.com/) capitalized on this global phenomenon, putting together a worldwide community of more than 15,000 users, collaborating in generating a nationwide genealogical dataset. The project began in 2010, from an idea of John Tierney developed by Maurizio Toscano, both still at the forefront of the initiative. It started and still runs as a community focused grassroots heritage project, where local community groups living in Ireland are trained in low-cost high-tech field survey of historic graveyards and recording of their own oral histories. In almost a decade, historicgraves.com published online more than 800 graveyards, recording at least a photo and the location of 99,235 graves and collaboratively transcribing the details of 199,851 people, and counting.

Community co-production happens within a freely available online platform, created for the transcription of memorial epitaphs. Training workshops are offered to local communities interested in contributing to surveying and transcribing historic graveyards. The combination of online interaction with local workshops and meetings works best in terms of ensuring meaningful participation. Data gathering and management procedures have proved to be essential on two fronts: data collected in the field is normally available online in short time; on the website, memorials' transcription is unmediated and immediately available to share. Instant publication proved to be highly engaging for the volunteer groups involved: they see immediate results for their work and are willing to share them with family and friends living abroad. Each local community can then download individual datasets of their own records, as tabular Open Data. The global community of users takes responsibility for quality control and completeness.

In 2018, the project has received a Heritage Council grant, that supported the upgrading of the online platform (backend and frontend), and it has been part of the initiative European Year of Cultural Heritage (#EuropeForCulture). From a technical point of view, data, code and layout have been upgraded from Drupal 6 to Drupal 7, with a plan to move to BackdropCMS or Drupal 9 in late 2021, when both Drupal 7 and Drupal 8 will reach "End of Life".

The Historic Graves project is fundamentally a Public History project, originated from local communities and that constantly relies on them to grow, solve ethical issues and stay sustainable. Technologies, and the Web in particular, have played a big role, both enabling remote transcriptions and stimulating participation, although they never been the focus of the initiative. Additionally, in line with public digital humanities principles, the project has always been open to participation, well beyond the academy, growing as a group of networked and engaged communities, not just as a repository for content.



References

Irish Abroad Unit (IAU) (2017) Irish emigration patterns and citizens abroad. 20 June. Available at: https://www.dfa.ie/media/dfa/alldfawebsitemedia/newspress/publications/ministersbrief-june2017/1-Global-Irish-in-Numbers.pdf (accessed 14/06/2020)Tierney, John, Kiely, Jacinta, & Toscano, Maurizio. (2019). The Historic Graves project. A Community Archaeology of Families

- A Call for Collaboration, in TEA. The European Archaeologist, ISSUE 60 - SPRING 2019, ISSN 1022-0135. Available at: https://www.e-a-a.org/EAA/Publications/Tea/Tea_60/Announcements/EAA/Navigation_Publications/Tea_60_content/Announcements.aspx#60_graves (accessed 14/06/2020)Toscano, Maurizio, & Tierney, John. (2016). Historicgraves.com: the online graves and graveyards finder. Zenodo. http://doi.org/10.5281/zenodo.3893694Toscano, Maurizio. (2019). Crowdsourcing Irish History. Zenodo. http://doi.org/10.5281/zenodo.3893803

Free Land for Whom? Mapping Land Acquisition and Dispossession under the Homestead Act, 1863-1912

Julius Wilm
wilm@ghi-dc.org
German Historical Institute Washington DC, United States of America Roy Rosenzweig Center for History and New Media, George Mason University, United States of America

Keywords: Homestead Act, Indian displacement, discrimination, GIS, real estate
Topics: English, North America, 19th Century, 20th Century, database creation, management, and analysis, spatial & spatio-temporal analysis, modeling and visualization, Geography and geo-humanities, History

The poster presents a geospatial database and dynamic web mapping project on homesteaders in the US West from 1863 to 1912. Mapping statistics on land claims (and their success and failure) at the level of local land districts in combination with complaints by unsuccessful claimants as well as military and congressional reports on squatting on Indian land, the project aims to create a birds eye view of the Homestead Act with a particular focus on the law's accessibility to people of color and it's entanglement with Indian displacement. The poster discusses the project's overall aims, its current status, as well as conceptual and technical challenges.

En France, on hurle – au Québec, on braille! Une comparaison diachronique de l'usage des verba dicendidans la littérature franco-québécoise et française standard

Svenja Simone Guhr
Guhr@linglit.tu-darmstadt.de
Technical University of Darmstadt, Germany

Keywords: Québec French, French, comparative literature and linguistics, verba dicendi, linguistic trends
Topics: Comparative (2 or more geographical areas), Europe, English, French, North America, 19th Century, 20th Century, electronic literature production and analysis, semantic analysis, Linguistics, Literary studies

Fafinez-vous si le chien jappe, le bébé braille et la lectrice se crochit de rire? La langue française au Québec se distingue par ses nombreuses particularités qui l'éloignent du français standard. Depuis leur séparation au 18ème siècle, les deux variétés françaises, en l'occurrence celle du Québec et celle de l'Hexagone, se sont développées indépendamment. Cependant, la littérature québécoise s'est enrichie d'un lexique marqué par des singularités linguistiques nombreuses comme des archaïsmes, des anglicismes, des néologismes et des régionalismes. De multiples recherches sociolinguistiques se sont attachées à relever les différences lexicales dans le vocabulaire du franco-québécois comparé au français dit standard international (cf. Meney 2010, 2017; Remysen et al. 2016).

Désormais, les nouvelles méthodes numériques offrent des opportunités innovantes qui permettent des analyses quantitatives. Les études linguistico-littéraires fondent désormais leurs recherches sur l'analyse de corpus numériques. Ces corpus conçus à partir d'œuvres littéraires formant ainsi une source de données textuelles constituent une base pour une analyse de fréquence et de la distribution du vocabulaire propre à une variété linguistique.

Le projet pilote ici décrit se trouve au carrefour de ces deux littératures: française et franco-québécoise. Il s'occupe des particularités et des différences résultant de l'évolution respective des deux variétés françaises qui se sont manifestées depuis leur séparation. En comparant les œuvres en prose du corpus, l'objectif de la recherche se focalise sur l'usage du groupe des verbes introducteurs dits verba dicendi (v.d.) utilisés pour introduire des discours directs et indirects dans des textes narratifs ainsi que pour marquer des interactions parlées ou des processus mentaux en général. L'analyse comparative dia- chronique vise à démontrer le développement du lexique et la signification sémantique des v.d. utilisés dans les œuvres littéraires écrites du 19ème siècle jusqu'au début du 20ème siècle en France et au Canada francophone.

À l'aide des méthodes quantitatives, on veut démontrer qu'il y a un rapport entre l'usage préféré d'une sélection des v.d. employés pour introduire des discours et l'origine de l'auteur(e) du texte ainsi que la présence des tendances d'un usage préféré de certains v.d. dépendant de la période temporelle et de la localisation du texte. La méthodologie suit un déroulement conventionnel: Après le relevé des textes appropriés pour l'analyse, un corpus a été élaboré qui comprend 125 œuvres francophones en prose de différent(e)s auteur(e)s dont 100 œuvres françaises de l'Hexagone (prises de la French Novel Collection(ELTeC-fra)) et 20 franco-québécoises publiées entre 1837 et 1937. Ainsi, le corpus a été reparti en trois périodes temporelles au service de la comparaison diachronique. Ensuite, les textes ont été annotés avec des informations linguistiques, telles que parties de discours et lemmes. Puis, les données ont été analysées avec l'objectif de calculer la fréquence de différents v.d. employés dans les textes soit de la variété franco-québécoise soit de celle de l'Hexagone. À la base d'un algorithme qui prend en compte les particularités morphologiques des v.d. et leurs caractéristiques syntactiques, les v.d. ont été extraits et cumulés pour ainsi recevoir une liste de fréquence individuelle pour chaque période temporelle considérée. En conséquence, ces listes serviront à une comparaison diachronique focalisée sur les tendances d'usage des v.d. dans les deux littératures. En somme, cette étude contribue aux recherches sémantiques historico-diatopique de ce groupe de verbes.

References

Bakhtin, Mikhail (1981): "Forms of Time and of the Chronotope in the Novel", in: The Dialogic Imagination, Texas: University of Texas Press.

Blodgett, Edward D. (2004): "Francophone writing", in: Kröller, Eva-Maria (ed.): The Cambridge Companion to Canadian Literature, Cambridge, UK: University Press 49-69.

Bloß, Alexandra (2005): Merkmale von Figurenrede in narrativen Texten: ein Vergleich zwischen deutschen und englischen Romanen und Kurzgeschichten, Ph.D. thesis, University of Wurzburg, DE.

Boivin, Aurélien (2015): Regards sur la littérature québécoise, Québec français 174, 65-68.

Brunner, Annelen (2015): Automatische Erkennung von Redewiedergabe: ein Beitrag zur quantitativen Narratologie, Berlin, DE: De Gruyter.

Chapman, Raymond (1984): The Treatment of Sounds in Language and Literature, Oxford, UK: Blackwell Publisher Limited.

Chapman, Rosemary (2013): What is Québécois Literature? Reflections on the Literary History of Francophone Writing in Canada, Liverpool, UK: University Press.

Gauvin, Karine (2011): L'Élargissement sémantique des mots issus du vocabulaire maritime dans les français Acadien et Québécois, Ph.D. thesis, University Laval, CA.

Genette, Gérard (1969): Figures. Essais, Paris, FR: Éditions du Seuil.

Groß, Konrad / Klooß, Wolfgang / Nischik, Reingard M. (eds.) (2005): Kanadische Literaturgeschichte, Stuttgart/Weimar, DE: Verlag J.B. Metzler.

Meney, Lionel (2017): Le français québécois entre réalité et idéologie, Un autre regard sur la langue, Étude sociolinguistique, Laval, CA: Presses de l'Université?.

Meney, Lionel (2010): Main basse sur la langue: idéologique et interventionnisme linguistique au Québec, Montréal, CA: Éditions Liber.

Meney, Lionel (1999): Dictionnaire québécois-français. Pour mieux se comprendre entre francophones, Montréal, CA: Guérin.

Pöll, Bernhard (20172): "Französisch außerhalb Frankreichs", in: Noll, Volker / Veldre-Gerner, Georgia (eds.): Romanistische Arbeitshefte, Band 42, Berlin/ Boston: De Gruyter 73-96.

Remysen, Wim (2016): "Langue et espace au Québec: les Québécois perçoivent-ils des accents régionaux?", in: Molinari, Chiara / Gavinelli, Dino: Espaces réels et imaginaires au Québec et en Acadie. Enjeux culturels, linguistiques et géographiques, Milano, IT: LED.

Schmid, Helmut (1994): "Probabilistic Part-of-Speech Tagging Using Decision Trees", in: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Schöch, Christof / Burnard, Lou (eds.) (2020): "French Novel Collection" (ELTeC-fra), Version v0.9.0-alpha, May 2020, in: European Literary Text Collection (ELTeC). COST Action Distant Reading for

European Literary History. DOI: https://doi.org/10.5281/zenodo.3462536.

Entry and Rise In Bureaucracy Patterns of Career Advancement in Joseon Dynasty

Donghyeok Choi
nerdinary@kaist.ac.kr
Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Korea, Republic of (South Korea)

Juyong Park
Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Korea, Republic of (South Korea)

Keywords: Quantitative History, Joseon Dynasty, Career Success
Topics: Asia, English, 15th-17th Century, 18th Century, 19th Century, network analysis and graphs theory and application, text mining and analysis, History, Humanities computing

Quantitative studies of social systems are thriving nowadays. These studies have paid particular attention to studies related to politics. It suggests that the political system is a constant interest, which is amply understandable. However, these studies have dealt with the byproducts of the actions and arguments by politicians, and as a result they do not examine the very operating dynamics of the political society itself, i.e. how the actors enter and fare inside it.

In this work, we analyze a comprehensive, well-preserved data covering a long time period in a stable political system, and tracks individual bureaucrats' career trajectories. It is the integrated data set from the Joseon Dynasty that ruled the Korean peninsula. The Annals of the Joseon Dynasty (AJD) and the Pangmok are the main components of our dataset. From these datasets, we could completed every career trajectory of 10,638 Joseon bureaucrats.

Using Bayesian Network Classifiers to Trace the Constuction of Race, Gender, and Class in all Print Editions of the Encyclopedia Britannica 1768-2010

Aaron Mauro
amauro@brocku.ca
Brock University, Canada

Keywords: machine learning, natural language processing, encyclopedia, reference, bayesian networks
Topics: Europe, English, North America, 18th Century, 19th Century, 20th Century, artificial intelligence and machine learning, natural language processing, African and African American Studies, Feminist studies

This short paper presentation will share recent discoveries made regarding conceptions of race, gender, and class across the 242 year history and all 15 editions of the Encyclopedia Britannica.

Extracting a Social Network of Musicologists

Thomas Nikolaus Haider
thomas.haider@ae.mpg.de
Max Planck Institute for Empirical Aesthetics, Frankfurt University of Stuttgart

Annette van Dyck-Hemming
annette.van-dyck-hemming@ae.mpg.de
Max Planck Institute for Empirical Aesthetics, Frankfurt

Keywords: Musicology, Named Entity Recognition, Social Network
Topics: Europe, English, 18th Century, 19th Century, 20th Century, artificial intelligence and machine

learning, text mining and analysis, Humanities computing, Musicology

0 Abstract

To aid musicological research, we extract named entities from a German musicological encyclopedia (MGG), and link these entities in a social network by the entities that they co-occur with. We offer two network plots that reveal details about the history of musicology, of (a) musicologists that have a lexicon entry each, and (b) the composers that are mentioned in these entries.

1 Introduction

Over the 18th and 19th century the academic discipline of musicology had become a central stage for negotiating the value of certain music and aesthetics. These negotiations have also led to the exclusion and deprecation of particular cultural elements, be they issues, music, or people. This discourse has been reflected in the encyclopedia 'Die Musik in Geschichte und Gegenwart' (Music in History and Present, MGG, comparable to the New Grove Dictionary of Music and Musicians). It contains 1282 biographical entries of musicologists, representing the Germany centered Western European discourse on musicological knowledge of the last 120 years.

To analyze this text corpus at scale, we train a Named Entity Recognition (NER) system and manually annotate texts from the encyclopedia to evaluate the NER on the new domain. We then employ a name standardization and cleanup procedure to finally link the entities and draw plots according to the properties of the graphs. Previous manual annotations showed that we might expect data that enables us to analyze the professional network of musicologists exposed in the MGG Online (van Dyck-Hemming and Wald-Fuhrmann, 2019). The approach in this paper already gives us an overview of social groupings, main actors and most negotiated composers (van Dyck-Hemming and Wald-Fuhrmann, 2016; Latour, 2014).

2 Named Entities

In order to obtain a NER system, we tune a pre-trained BERT model (Devlin et al., 2018) on the CoNLL-2003 Named Entity Dataset (Sang and De Meulder, 2003) with the deepset.ai toolkit and achieve competitive performance on 10% held out data. See Table 1 for an overview of the F1-scores of our NER model on CoNLL-2003 itself and Table 2 for an evaluation on manual annotation of students.

	precision	recall	f1-score	support
PER	0.92	0.92	0.92	1693
LOC	0.91	0.92	0.91	2376
ORG	0.80	0.81	0.81	1330
OTH	0.70	0.66	0.68	778
micro avg	0.84	0.87	0.85	6177
macro avg	0.86	0.87	0.86	6177

Table 1: Named Entity Recognition Evaluation on held out CoNLL-2003 data

	precision	recall	f1-score	support
PER	0.44	0.62	0.52	593
LOC	0.71	0.51	0.59	624
ORG	0.22	0.29	0.25	452
OTH	-	-	-	-
micro avg	0.24	0.49	0.32	1669
macro avg	0.48	0.49	0.47	1669

Table 2: Named Entity Recognition Evaluation of Manual Annotation on MGG

Two students annotated around 3000 tokens with the classes Person (PER), Geopolitical Entity/Location (GPE/LOC), Organizations (ORG), other (OTH), temporal expressions (TIME), and professional field/discipline (FACH). We calculated Cohen's kappa between the annotators on document level. Over all classes, agreement ranges between .5 and .8. When removing FACH, agreement increases to .7 to .95, suggesting that future research should implement proper guidelines for this label. Our best model achieves a F1-score of 85% on CoNLL-2003, but on the manual annotation (with TIME, FACH, OTH removed), it only achieves 47% F1-macro. Especially ORG is not reliably detected. Since manual inspection suggested good performance of the model for the other classes, we suspect that performance would benefit from (a) paying special attention to punctuation marks (commas, brackets) that are ubiquitous in encyclopedic articles (b) improved training of the annotators and tailoring the guidelines to the used model, and (c) domain adaptation of BERT. As recall for persons and locations is acceptable, we perform additional clean-up and standardization of names.

2.2 Name Standardization

For persons, we automatically remove all first names and thus only use surnames. See Figure 1 for an overview of the cleaning pipeline. Unfortunately, only relying on last names leads to the conflation of some names, e.g., Hermann and Amalie Abert, the latter being the first female academic musicologist in Germany.

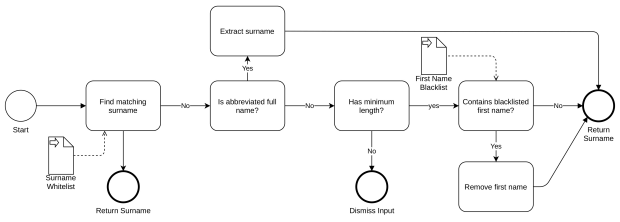


Figure 1: Name Standardization Process

3 Social Network

To construct a social network, we select a subset of names (either musicologists only, or musicologists and composers) as nodes and draw an undirected edge (connect) every name in a particular encyclopedia article only to the musicologist in the title of the

article. The network visualization is performed with the software Gephi (Bastian et al., 2009). We focus on the Largest Connected Component (Newman, 2010) and discard nodes with a degree of less than three. The Modularity measure (Blondel et al., 2008) is used to color-encode modules within the network, while the Eigenvector Centrality (Newman, 2010) embraces nodes of higher importance by increasing the node (font) size to scale. Finally, we apply the OpenOrd (Martin et al., 2011) or ForceAtlas2 (Jacomyet al., 2014) layout algorithms to draw the final graphics. In addition to a layout, these algorithms produce a visual clustering that can give further insights.

3.1 Social Network of Influential Actors in Musicology

Figure 2 illustrates a graph of musicologists that are connected whenever they occur in each others biography, and select only names that also occur in the title of an encyclopedic article. We find a central group of researchers that minted German musicology. This group spans from Hermann and Amalie Abert over Arnold Schering, to Hugo Riemann and Heinrich Kretzschmar. The likewise important Erich Hornbostel and Curt Sachs represent the field of systematic musicology. Other researchers sharing their interests are grouped around them. Guido Adler, who is loosely connected to many names, likely because he founded the first institute of musicology in Vienna, drifted off center to house his students that are also connected to the center. We can even partly validate a study based on quantified historical data on Carl Dahlhaus and Hans Heinrich Eggebrecht who represent the two main figures of German musicology from 1965 to 1995. We observe a clear distinction between both researchers and are able to identify a significantly bigger group connected to the younger Dahlhaus.

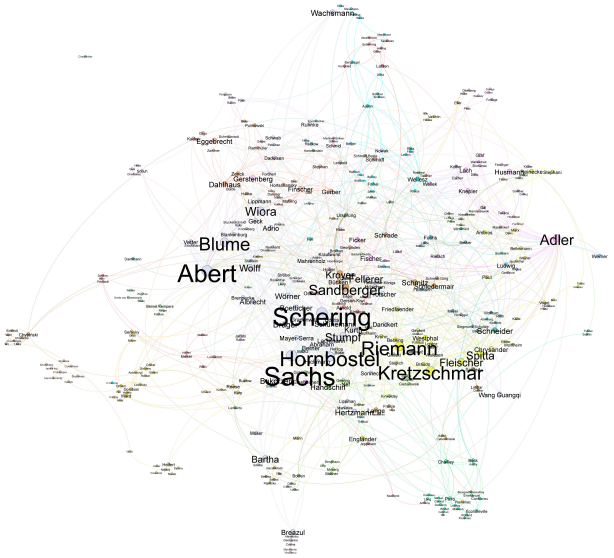


Figure 2: Social Network of Musicologists

3.2 Musicology and its Issue

In addition to the musicologists, the data basis of Figure 3 also includes all names of composers. Here, the visualization shows a network of musicology and its main issues (as composers are only connected to the musicologists, but not directly to each other). Composers from the Bach-family and Mozart are the issues of musicology as an academic discipline. So, the overwhelming dominance of names like 'Mozart', 'Bach', 'Beethoven' and 'Wagner' represent the dominance of the research that musicologists have done on these composers and their music. Especially Bach resp. the Bach-family seems to be the one issue deeply connected to the earliest and most important musicologists namely Johann Forkel, Philip Spitta, and Guido Adler. Moreover, there are numerous edges (identifiable via color) linking this group with younger musicologists like Arnold Schering, Hermann Abert, Friedrich Blume and others who shaped musicology of the 1920s.

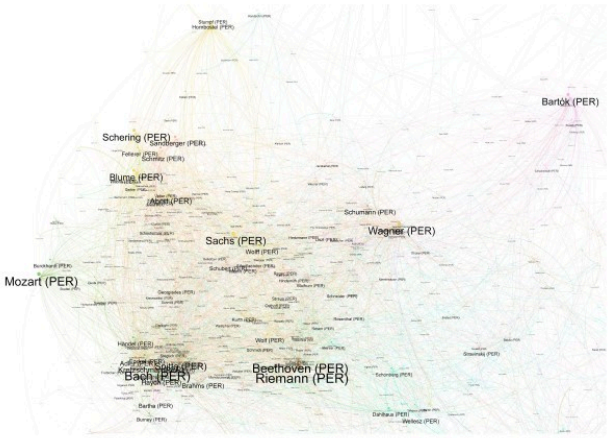


Figure 3: Network of Musicologists and Composers

4 Conclusion

We presented a first study of a clearly defined corpus of prosopographical texts, from which we extracted Named Entities and linked these to visualize the social network of musicologists and their relations to certain issues: the composers. This synchronic approach is already useful for historical research to confirm hypotheses that were previously drawn from arduous manual work. Regarding musicology, we could confirm quantitatively what we expected and already knew from conventional historical research. The overwhelming dominance of a few composers also allows us to reflect about the knowledge standards that musicology has so far relied upon.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10):P10008.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PLOS ONE, 9(6):1–12.

Latour, B. (2014). Reassembling the social—An Introduction to Actor-Network-Theory. Oxford University Press, 3 edition.

Martin, S., Brown, W., Klavans, R., and Boyack, K. (2011). Openord: An open-source toolbox for large graph layout. Proc SPIE, 7868:786806.

Newman, M. (2010). Networks: An Introduction. Oxford University Press, Inc., USA.

Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.

van Dyck-Hemming, A. and Wald-Fuhrmann, M. (2016). Vom datum zum historischen zusammenhang: M¨oglichkeiten und grenzen einer fachgeschichtlichen datenbank. In Wissenskulturen der Musikwissenschaft. Generationen–Netzwerke–Denkstrukturen, pages 261–278. transcript.

van Dyck-Hemming, A. And Wald-Fuhrmann, M. (2019). Ansätze zur analyse sozialer netzwerke in der musikwissenschaft mit hilfe von graphentechnologie. Poster presented at the Jahrestagung der Gesellschaft f¨ur Musikforschung, university of Paderborn/Detmold 2019.

Construction of the corpus of senmyō: one of the oldest materials of Japanese language

Toshinobu OGISO
togiso@ninjal.ac.jp
NINJAL, Japan

Neisin GO
gons@ninjal.ac.jp
NINJAL, Japan

Yukie IKEDA
yu.ikeda71@gmail.com
Chuo University, Japan

Tetsuya SUNAGA
tetsuyasunaga@gmail.com
Showa Women's University, Japan

Keywords: History of Japanese language, Japanese Linguistics, Corpus linguistics
Topics: Asia, English, 5th-14th Century, manuscripts description, representation, and analysis, natural language processing, Asian studies, Linguistics

We worked on construction of the corpus of senmyō (imperial edict) written in the 8th century for linguistic research. Senmyō is one of the oldest materials of Japanese language and is written in Old Japanese

using a special notation method using only Chinese characters called "senmyō-gaki". In order to encode this notation, we reproduced it using a originally extended tag set based on TEI. We also added word information to the full text of this corpus using Mecab and UniDic. As some of the words in senmyō can be read in two ways, Chinese style and Japanese style, we devised that these two readings can be assigned to the same location when adding word information. This corpus is published through an online search application called "Chunagon".

Acquisition and Analysis of a Meme Corpus to Investigate Web Culture

Thomas Schmidt
University of Regensburg, Germany

Philipp Hartl
University of Regensburg, Germany

Dominik Ramsauer
University of Regensburg, Germany

Thomas Fischer
University of Regensburg, Germany

Andreas Hilzenthaler
University of Regensburg, Germany

Christian Wolff
christian.wolff@ur.de
University of Regensburg, Germany

Keywords: Memes, Web Culture, Text Mining, Sentiment Analysis, Topic Modeling
Topics: Europe, English, Contemporary, social media analysis and methods, text mining and analysis, Cultural studies, Media studies

1. Introduction

Memes are a popular part of today's online culture, reflecting current developments in pop culture, politics or sports. That has led various scholars in the humanities and other research areas to examine the importance and role of memes (Shifman, 2014a; Highfield & Leaver, 2016; McCulloch, 2019). Bauckhage (2011) defines the term Meme as "contents or concepts that spread rapidly among Internet users". While memes with solely visual content are rising in popularity, one of the most common and historically important meme types is the "image macro" which consists of a reusable image template with a top and/or bottom text (figure 1).

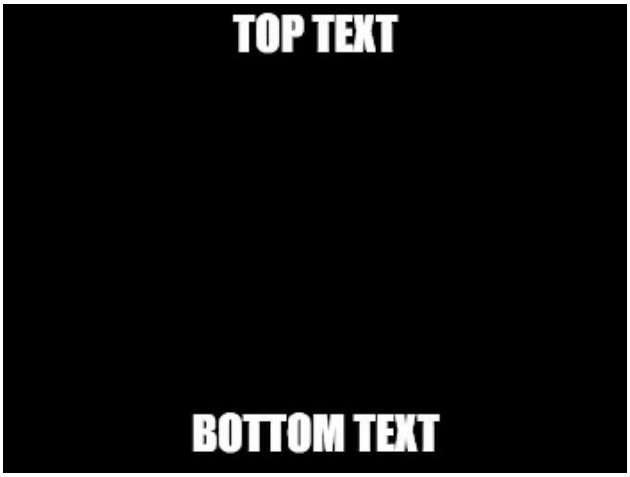


Figure 1: Typical format of an image macro

There are various established image templates (see figure 2 for an example) and with the growth of social media, new ones are constantly emerging. We differentiate between the meme template, which is basically just the image of a meme and the meme derivatives, which are the multiple manifestations of a meme template differing regarding the text of the meme.

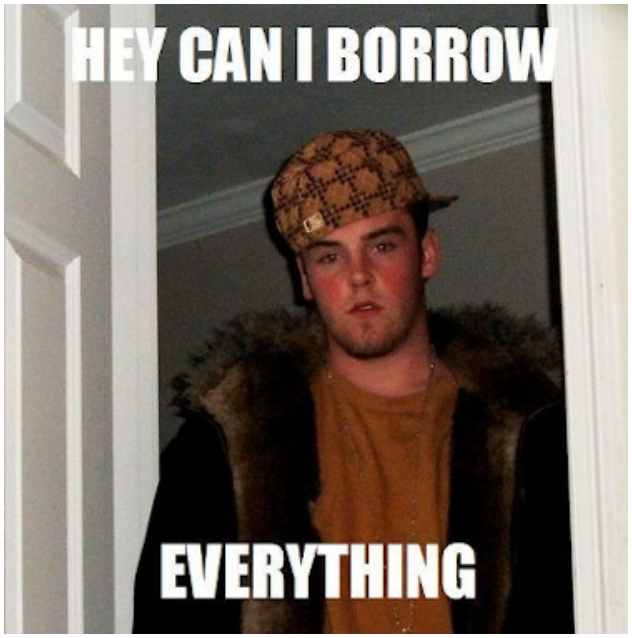


Figure 2: Example of "Scumbag Steve", a popular image macro meme

Although memes are distributed and shared in large quantities, the majority of current research on memes is qualitative, e.g. analyzing patterns and stylistic rules of a small number of memes (Shifman, 2014b; Osterroth, 2015). Since image macros typically have a textual component, we want to use computational methods of Distant Reading (Moretti, 2013) to analyze memes in a large-scale approach. Our project aims to identify developments of the content and sentiment of memes in a diachronic way but is also based on image templates. In this paper we present first results on the corpus acquisition workflow we have developed as well as the application of general text analysis, topic modeling, and sentiment analysis on the overall corpus.

2. Corpus Creation

To create a corpus for our analysis we use the platform knowyourmeme (https://knowyourmeme.com/). It is one of the most popular platforms for uploading memes and offers the possibility to search for specific meme categories like image macros. Furthermore, the different derivatives of a meme template are collected under a single entry and are enriched with metadata. For our first analysis, we focus on 16 of the historically most popular templates and we have implemented a scraper to access the links to the meme derivatives and metadata. To get the text of the memes we use Google Cloud OCR on the images gathered. Our final dataset consists of 7.840 meme derivatives, metadata and the text (see figure 3). This corpus is publicly available for the research community to download and use: https://docs.google.com/spreadsheets/d/1ESecAyzfWvXTELflbWwZEXWE_tGKq8wmTTpOTSSrDJY/edit?usp=sharing

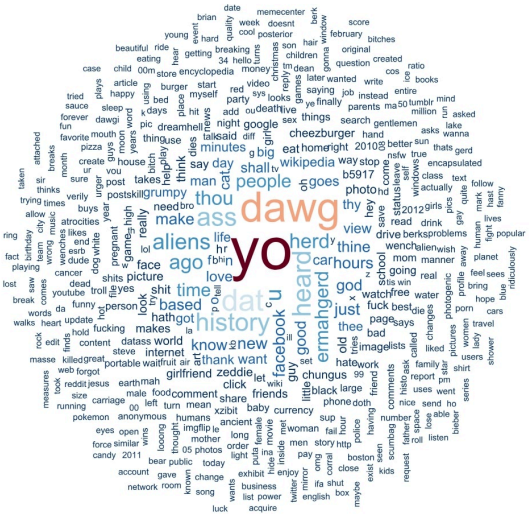
Please note that we only include memes with English language since this is the language knowyourmeme is focused on.

#	Template	Amount of macros	Total tokens	Average tokens per macro
1	'Dat Ass	403	2122	5
2	Based God	129	950	7
3	[10] Guy	213	1930	9
4	Ancient Aliens	639	5494	9
5	Bad Luck Brian	674	7411	11
6	Ermahgerd	330	3538	11
7	Grumpy Cat	509	5654	11
8	Philosoraptor	649	7923	12
9	Scumbag Steve	474	5658	12
10	Ridiculously Photogenic Guy	322	4370	14
11	Annoying Facebook Girl	234	4056	17
12	Joseph Ducreux / Archaic Rap	930	16980	18
13	First World Problems	346	6584	19
14	Overly Attached Girlfriend	205	3862	19
15	Big Chungus	111	2661	24
16	Xzibit Yo Dawg	629	20175	32

Figure 3: Corpus description

3. Corpus Analysis

For all approaches, we have implemented various preprocessing steps commonly used in text mining (e.g. lemmatization). Figure 4 shows a word cloud of the most frequent words of the entire corpus:


yo
dawg
ermahgerd
aliens
ass
make
thou
people
herp
derp
thine
view
ago
face
love
time
based
history
facebook
know
new
thank
want
girlfriend
zeddie
chungus
hate
work
just
god
good
guy
friends
share
comment
food
black
phone
currency
baby
xzibit
ancient
humans
anonymous
pokemon
eyes
open
wins
force
similar
candy
2011
bear
public
today
account
gave
network
room
song
wants
business
list
power
luck
acquire
exhibit
enjoy
twitter
ifa
shut
box
maybe
english
mirror
close
kids
exist
seen
scumbag
father
request
pay
boston
men
story
http
woman
fail
police
hour
having
sup
number
nice
send
ho
able
bieber
tote
listen
pm
report
family
shirt
series
comments
hug
ask
called
liked
pictures
women
users
travel
shower
lady
uses
went
changes
porn
cars
photogenic
profile
away
planet
bring
hope
blue
ridiculously
lives
popular
human
light
hate
funny
mark
quite
pics
gay
girls
2012
drive
berks
problems
wench
alien
wish
real
sees
feel
going
mom
manner
school
watch
free
water
page
says
bad
thee
liked
changes
images
lists
die
best
fuck
porn
cars
hope
bring
planet
away
profile
photogenic
pictures
women
users
travel
shower
lady
uses
went
series
shirt
family
pm
report
ask
hug
comments
friend
doth
send
ho
able
bieber
tote
listen
father
scumbag
seen
kids
exist
request
omg
ifa
shut
box
maybe
english
mirror
close
boston
pay
men
story
http
woman
fail
police
hour
having
sup
number
nice

For topic modeling, we use Latent Dirichlet Allocation (LDA, Blei et al., 2003) to calculate 16 LDA topics. LDA topics are described by typical word clusters within documents (here: meme derivatives), thus topic modeling produces lists of words that appear frequently together in documents. Our assumption is that every meme template is equivalent to a topic, thus we chose the number of our image macros as topic number (16).

Figure 5 illustrates our results for the topic modeling analysis:

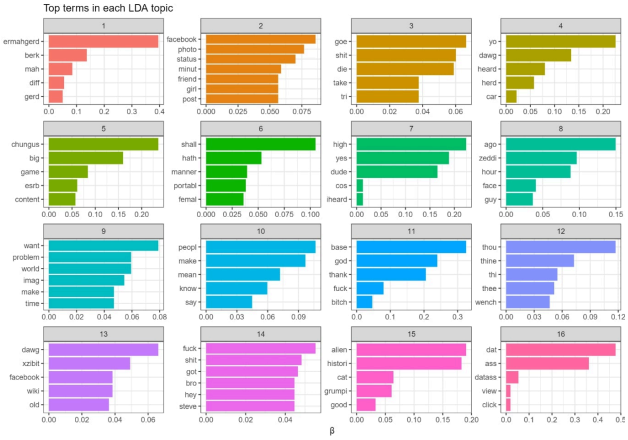


Figure 5: 16 LDA topics of the corpus; with the five most contributing tokens per topic

As expected, most of the topics are expressions of a single meme template (e.g. topic 1 for the "Ermahgered" or topic 3 for the "XZibit Yo Dawg" meme template) which shows that some memes consist of homogenous and reoccurring word patterns. However, there are some overlaps like topic 15, expressing words common in the "Ancient Alien" and "Grumpy Cat" meme. We plan to investigate these memes in future work in more detail to examine the similarities they have in more detail. For the sentiment analysis, we use the sentiment lexicon "Bing" (Liu, 2012; Liu & Zhang, 2012) for polarity (positive, negative) and the NRC Word-Emotion Association Lexicon (Mohammad & Turney, 2013) for emotions. Figure 6 shows which words contribute the most to a specific overall sentiment:

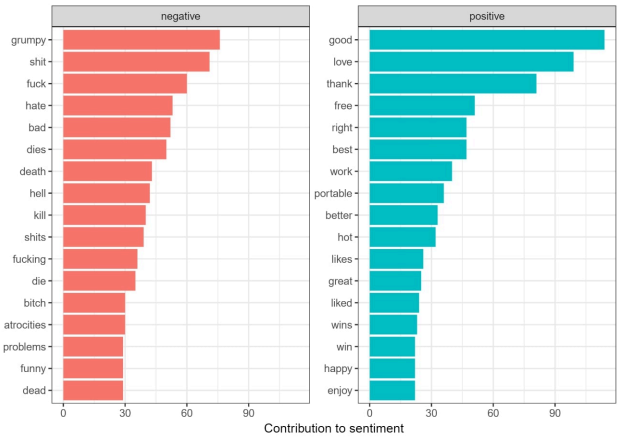


Figure 6: Most important tokens contributing to the overall sentiment in the corpus

Though we cannot report the results of the sentiment and emotion comparisons among the memes in detail, one outlier meme we want to highlight is the "Ancient Alien" meme. The "Ancient Alien" meme has the highest values for disgust and fear, which is a fitting result since those memes are often used in the context of conspiracy theories.

Currently, our research is at an early stage and exploratory. In future work, we want to continue our analysis by increasing our corpus, filtering out noise during the acquisition and gather more metadata to perform diachronic and meme based analysis and comparisons considering sentiments and topics.

References

Bauckhage, C. (2011, July). Insights into internet memes. In Fifth International AAAI Conference on Weblogs and Social Media.

Bauckhage, C., Kersting, K., & Hadiji, F. (2013, June). Mathematical models of fads explain the temporal dynamics of internet memes. In Seventh International AAAI Conference on Weblogs and Social Media.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Davison, P. (2012). The language of internet memes. The social media reader, 120-134.

Highfield, T., & Leaver, T. (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. Communication Research and Practice, 2(1), 47-62.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining text data (pp. 415-463). Springer, Boston, MA.

McCulloch, G. (2019). Because Internet: Understanding the new rules of language. Riverhead Books.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465.

Moretti, F. (2013). Distant reading. Verso Books.

Osterroth, A. (2015). Das Internet-Meme als Sprache-Bild-Text. Image, 22, 26-46.

Shifman, L. (2014a). Memes in digital culture. MIT press.

Shifman, L. (2014b). The cultural logic of photo-based meme genres. Journal of Visual Culture, 13(3), 340-358.

Building Bridges: Pedagogical Strategies for Introducing Digital Humanities in the Undergraduate and Graduate Classroom

Sarah Ketchley
ketchley@uw.edu
University of Washington, United States of America

Wendy Perla Kurtz
wpkurtz@ucla.edu
UCLA, United States of America

Keywords: pedagogy, collaboration, digital literacy
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, Education/ pedagogy

This poster will highlight some of the challenges of teaching introductory-level digital humanities courses in undergraduate and graduate classrooms, and will describe pedagogical solutions developed by faculty at UCLA and the University of Washington to address these problems. These include identifying and developing the core skill sets students need to begin work in digital humanities, such as best practices for project management, working with data and interpreting and presenting analyses. From a faculty standpoint, the poster will suggest strategies for building collaborative partnerships between libraries and faculty to best leverage each respective group's expertise.

Not Coding Alone: Teaching the Collaborative Habits of DH in a MOOC

Stephen Osadetz
osadetz@fas.harvard.edu
Harvard University, United States of America

Cole Crawford
cole_crawford@fas.harvard.edu
Harvard University, United States of America

Christine Fernsebner Eslao
eslao@fas.harvard.edu
Harvard University, United States of America

Keywords: pedagogy, teaching, collaboration, mooc, online courses
Topics: Global, English, Contemporary, public humanities collaborations and methods, text mining and analysis, Education/ pedagogy, Humanities computing

Introduction

Online courses present special challenges to the teaching of the digital humanities, particularly having to do with the collaborative nature of the discipline. In the last decade, as digital humanities pedagogy has emerged as a coherent and visible discipline in its own right (Hirsch 2012), the requirement that "students need to develop their collaborative and interdisciplinary skills" has arisen as a central concern (Mahony and Pierazzo, 215).

This poster describes the collaborative and project-based approach of our course, Digital Humanities in Practice, a forthcoming HarvardX MOOC (June 2020). Our aim is to embody the pedagogic ethic of "seamfulness" advocated by Mauro et al. in a recent special issue of DHQ, productively exposing the messy process of development. We do so in a number of different ways: 1) the course is team-taught by a scholar, a technologist, and a librarian to demonstrate the affordances and difficulties that arise when the instructors' different competencies are knit together; 2) we encourage our students to make mistakes and we

model doing so ourselves, so they can see how the progress of a digital project often does not follow a straight line, and 3) while the arc of the course draws the complete narrative of a digital project -- from corpus acquisition to publication and project maintenance -- the pedagogical patterns of the course reinforce the iterative nature of digital work.

Project Background

In his own research on eighteenth-century intellectual history, Stephen Osadetz assembled a team of computer technologists and librarians that created Open Books (conceptsearch.fas.harvard.edu), a concept-search engine for Gale's Eighteenth Century Collections Online database. Based on the success of this project, Osadetz and his team were asked to develop a practical course focused on text mining that embodies the collaborative nature of their work.

Our team held a number of focus groups with students who had enrolled in the original course, and Cole Crawford conducted a survey of over 100 DH graduate certificate and degree-granting programs. The results indicated a substantial growth in residential DH programs at all levels over the past five years, but an almost complete absence of online DH instruction.

Pedagogical Method

The course walks students through each part of a text analysis project, from acquiring data through presenting results. Units follow a set rhythm, which is meant to ease students into each new skill. After a brief assignment and video to frame what is to be taught, each unit breaks into two substantial parts: one guided, in which students follow along with an instructional video in whatever platform we are using (Jupyter Notebook, OpenRefine, Tableau, etc.); and the other self-guided, in which they practice and extend what they have learned, individually and with peers through a discussion board. These second, more exploratory assignments (on, for example, metadata enrichment with Wikidata, or author attribution, or LDA) will be peer-assessed. A preliminary study suggests that peer grading and professors' grading is highly correlated (Kaplan et al., 2014).

Conclusion

We have attempted to turn these difficulties of creating a comprehensive course on digital methods to our advantage by designing a MOOC that embraces the ethics of seamfulness and collaboration, addressing itself to different types of audiences: scholars embarking on digital research projects, but also technologists and librarians who increasingly facilitate this work. We believe this course may satisfy a crucial need that may be appreciated by many people interested in DH but intimidated by coding: a playful but intellectually challenging introduction to the field that tells students, above all, that you don't have to do everything on your own.

References

Deegan, Marilyn and Willard McCarty, editors. Collaborative Research in the Digital Humanities. Farnham, Surrey, England; Burlington, VT: Ashgate, 2012.

Hirsch, Brett D., editor. Digital Humanities Pedagogy: Practices, Principles, and Politics. Cambridge: Open Book Publishers, 2012.

Terdfras, Melissa, Deb Verhoeven, Claire Clivaz, Frédéric Kaplan, and Cyril Bornet. A Preparatory Analysis of Peer-Grading for a Digital Humanities MOOC, Digital Humanities 2014. Lausanne, Switzerland, 2014. https://infoscience.epfl.ch/record/200911.

Mauro, Aaron et al. "Towards a Seamful Design of Networked Knowledge: Practical Pedagogies in Collaborative Teams." Imagining the DH Undergraduate, special issue of DHQ, vol. 11, no. 3, 2017. http://www.digitalhumanities.org/dhq/vol/11/3/000322/000322.html

Mohoney, Simon and Elena Pierazzo, "Teaching Skills or Teaching Methodology?" Digital Humanities Pedagogy: Practices, Principles, and Politics. Brett D. Hirsch, ed. Cambridge: Open Book Publishers, 2012.

Murphy, Emily Christina and Shannon R. Smith. "Introduction." Imagining the DH Undergraduate, special issue of DHQ, vol. 11, no. 3, 2017. http://www.digitalhumanities.org/dhq/vol/11/3/000334/000334.html

A Priority Programme on Computational Literary Studies

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

José Calvo
University of Würzburg, Germany

Nils Reiter
University of Stuttgart, Germany

Evelyn Gius
University of Darmstadt, Germany

Christof Schöch
University of Trier, Germany

Simone Winko
University of Göttingen, Germany

Jonas Kuhn
University of Stuttgart, Germany

Fotis Jannidis
University of Würzburg, Germany

Keywords: literary studies, projects, text mining
Topics: Europe, English, 18th Century, 19th Century, 20th Century, semantic analysis, text mining and analysis, Literary studies

Beginning in 2020, the German Research Foundation (DFG) is funding projects within the framework of a new priority programme on "Computational Literary Studies" (CLS). The funding agency is financing ten individual research projects in Germany and Switzerland.

The research projects focus on mostly German literary fiction - including novels, drama and poetry - and encompass a broad spectrum of computational methods including network and sentiment analysis, stylometry and natural language processing, with the explicit aim of advancing new methods in the research field. The projects are highly interdisciplinary, in most of them digital humanists working in close cooperation with literary scholars on questions highly relevant for the current discourses in literary studies.

The poster informs about the programme and its constituent projects with the aim to reach out to the international community, and encourage researchers and organizations to connect, network and discover opportunities for cooperation and exchange.

Tracking the digital dead: A progress report in the use of transfer learning, image analysis, and forensic anthropology to classify human remains sold on social media

Damien Huffer
Carleton University, Ottawa

Shawn Graham
ShawnGraham@cunet.carleton.ca
Carleton University, Ottawa

Keywords: illicit trafficking, human remains trade, forensic anthropology, social media
Topics: Global, English, Contemporary, artificial intelligence and machine learning, mobile applications development and analysis, Anthropology, Media studies

This short presentation presents preliminary results of new work expanding on that of the ongoing Bone Trade Project (https://bonetrade.github.io/). It discusses progress in the use of transfer learning to create a hand-held application to classify human remains observed for sale online. Global legal loopholes allow the trade, sale, or auction of human remains of questionable provenance and provenience, creating the need to better map trafficking across platforms and understand its socio-cultural context as a community of practice. Uniting osteology, forensic anthropology and image analysis of large corpuses from public Facebook, Instagram, etc., this presentation demonstrates the potential to monitor this trade in real time. A progress report of mobile app development will be presented, as well as ongoing and future technical and ethical issues that have surfaced to date. Ultimately, we demonstrate how the technologies of surveillance can be subverted to restore some measure of humanity to the commodified dead.

The Hierarchy of Characters: Quantitative Dominance Relations in Russian Drama

Eugenia Ustinova
esustinova@edu.hse.ru
National Research University Higher School of Economics

Keywords: Network analysis, drama, digital literary studies
Topics: Global, English, 18th Century, 19th Century, 20th Century, network analysis and graphs theory and application, Literary studies

Network analysis methods are not a novelty in literary studies and in drama's research in particular. However, one approach is not enough to cover all the nuances of characters' relations. In this study, I combined network analysis methods and quantitative measures. This work is the first look at 'quantitative dominance relations' in Russian drama. Also, I present the method to divide characters of each play into groups of importance. It is used to assess the development of the structure of dramatis personae over time.

An interactive multimedia companion to Wagner's Lohengrin encoding and visualising a motivic study

Kevin Page
kevin.page@oerc.ox.ac.uk
Oxford e-Research Centre, University of Oxford, United Kingdom

David Lewis
david.lewis@oerc.ox.ac.uk
Oxford e-Research Centre, University of Oxford, United Kingdom

Laurence Dreyfus
Faculty of Music, University of Oxford, United Kingdom

Keywords: digital musicology, linked data, music encoding, multimedia
Topics: Europe, English, 19th Century, linked (open) data, music and sound digitization, encoding, and analysis, Humanities computing, Musicology

We introduce a tablet based interactive application which presents a comprehensive digital exploration as a companion to a musicological article. Taking the operaLohengrin , the article and digital companion show how one motive is altered each time it recurs, reflecting its role in the drama, and exploring Wagner's sophisticated treatment of recurring themes. Musicological analysis is encoded using Linked Data as an independent, repurposable, and open Research Object. Interactive user views are generated directly and dynamically in the browser from this knowledge graph using novel visualisations, which in turn enable the user to navigate all possible paths through the evidential multimodal materials. Our companion explores the different compositional devices Wagner uses to vary his motives, browsing the whole opera for motive occurrences and their musical and textual contexts. Visualisations and recordings support the analysis, making it accessible to an audience that may struggle with a Wagnerian orchestral score.

Musicological argument has traditionally been communicated in writings that are textual, linear and illustrated by occasional figures, despite almost always being concerned with diverse subjects and evidential materials, each potentially exemplified by different media, and each potentially a springing point for digression and exploration of the author's argument. While exploring the referenced materials may be non-linear, this is neither embodied in nor enabled by traditional communication mediums.

Previously we have digitally enhanced an extract of a musicological article[1] with dynamic and interactive elements. Here, we introduce a tablet based interactive application which presents a comprehensive digital exploration as a companion to a complete musicological article. Taking the opera Lohengrin, the article and digital companion show how one motive is altered each time it recurs, reflecting its role in the drama, and exploring Wagner's sophisticated treatment of recurring themes.

Musicological analysis is encoded – along with relationships to multimedia materials – using Linked Data as an independent, repurposable, and open Research Object. Interactive user views are generated directly and dynamically in the browser from this

knowledge graph using novel visualisations, which in turn enable the user to navigate all possible paths through the evidential multimodal materials.

Our companion explores the different compositional devices Wagner uses to vary his motives, browsing the whole opera for motive occurrences and their musical and textual contexts. Visualisations and recordings support the analysis, making it accessible to an audience that may struggle with a Wagnerian orchestral score. Exploration of this material can follow or be triggered by the article (Fig.1), but can also be reader-driven, with free browsing of the curated musical landscape. A video essay also provides a source of narrative paths through the companion, as a guide itself and as a source of starting points.

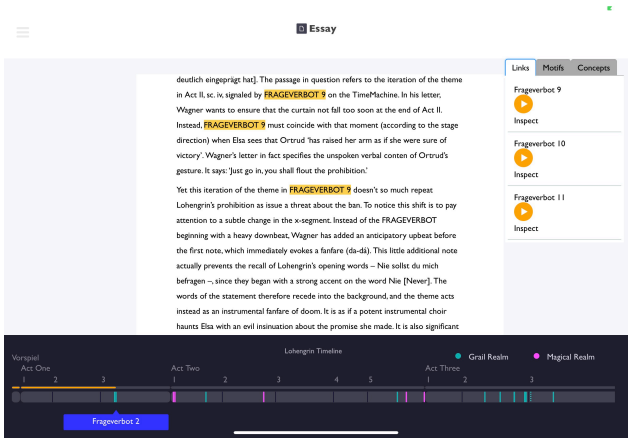


Figure 1: The musicological essay view in the digital companion.

The application is built with a new version of the MELD (Music Encoding and Linked Data) framework[2]. MELD traverses Linked Data graphs to select and filter relevant information, with reusable components for creating and retrieving annotations, and for displaying and interacting with musical, textual, graphical and audio-visual materials. MELD is written in Javascript and Python, with resources using standards including the MEI[3] music encoding, TEI, the Music Ontology and Web Annotations[4].

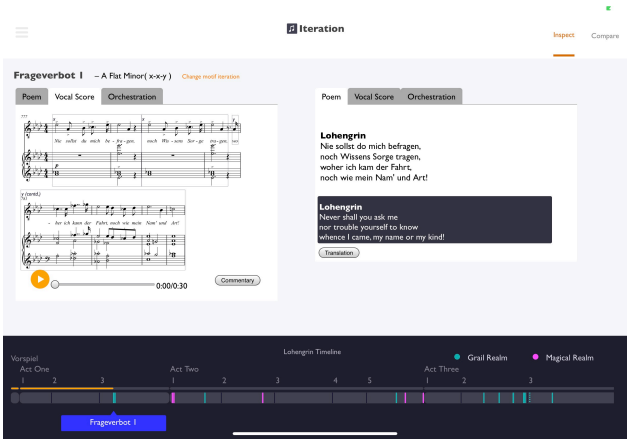


Figure 2: A view with labelled vocal score, text and translation.

We provide two views for music notational content. Vocal score reductions are rendered from MEI with structural analysis dynamically overlaid (Fig.2); annotations trigger audio playback from that point. A second notational visualisation of MEI simplifies the complexity of a Wagnerian orchestral score: each instrument playing at a particular time is shown as a coloured ribbon, with the instrument's section of the orchestra providing the colour (Fig.3). This highlights differences across orchestration iterations that may seem identical as vocal score.

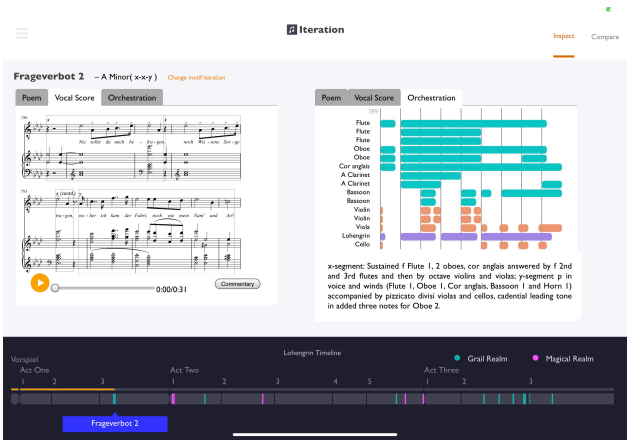


Figure 3. The same iteration showing the orchestral summary view.

For an opera thousands of bars long, overviews are crucial. An ever-present timeline shows all occurrences of a motive, providing a visual summary and a base for navigation. In the Time Machine view (Fig.4), users can also flick through motive occurrences – visualised as libretto, vocal score or orchestration –

summarising the sequence within the opera, supporting quick comparisons, and as an index to detail views.

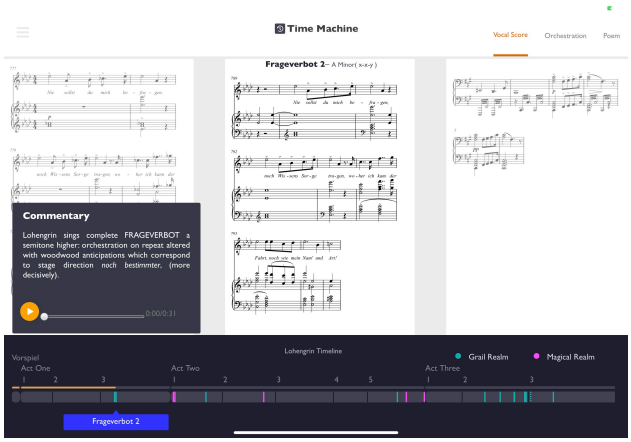


Figure 4. The Time Machine view in the Digital Companion.

Links to software and sourcecode:
The Lohengrin TimeMachine Digital Companion app and introductory video: https://um.web.ox.ac.uk/lohengrin Lohhengrin app source code: https://github.com/oerc-music/ForbiddenQuestion

The MELD framework the app is build upon: https://github.com/oerc-music/meld

MELD components in turn use open source libraries including:

Verovio: https://github.com/rism-ch/verovio

CETEIcean: https://github.com/TEIC/CETEIcean

[1] D. Lewis, D.M. Weigl, J. Bullivant, and Kevin R. Page. Publishing musicology using multimedia digital libraries: Creating interactive articles through a framework for linked data and mei. In Proc. DLfM, DLfM '18, pages 21–25, New York, NY, USA, 2018. ACM.

[2] D.M. Weigl and K.R. Page. A Framework for Distributed Semantic Annotation of Musical Score: "Take It To The Bridge!". In Proc. ISMIR 2017, Suzhou, China, October 23-27, 2017, pages 221–228, 2017.

[3] T. Crawford and R. Lewis. Review: Music encoding initiative. Journal of the American Musicological Society, 69(1):273–285, Spring 2016.

[4] M.E. Barnes, N. Ledchumykanthan, K. Pham, and K. Stapelfeldt. Annotation-based enrichment of digital objects using open-source frameworks. Code4Lib Journal, (37), 2017.

Computer Vision Approaches to Body Language and Facial Expressions Analysis in XX-century Cinema

Alena Koroleva
ITMO University, St. Petersburg, Russia

Danila Bulavkin
Tomsk State University, Tomsk, Russia

Evgenii Glazunov
National Research University Higher School of Economics, Moscow, Russia

Ildar Belyalov
ibelyalov@yandex.ru
National Research University Higher School of Economics, St. Petersburg, Russia

Mariia Knysheva
National Research University Higher School of Economics, Moscow, Russia

Nikita Severin
Southern Federal University, Rostov-on-Don, Russia

Oleg Serikov
National Research University Higher School of Economics, Moscow, Russia

Keywords: Cinematography, Expressiveness, Gestures, Computer Vision
Topics: Comparative (2 or more geographical areas), Global, English, 20th Century, Contemporary, artificial intelligence and machine learning, image processing and analysis, Film and cinema arts studies

The research aims to detect the difference between silent/sound movies and analyze the transitio n period from the perspective of movements. We use the computer vision tools to compute the needed characteristics of movies and perform statistical analysis to test our hypotheses.

A simple method of extracting keywords from texts

Maciej Eder
maciejeder@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences), Poland Pedagogical University of Krakow, Poland

Michał Woźniak
Institute of Polish Language (Polish Academy of Sciences), Poland

Keywords: keywords analysis, keywords, Zeta, TF-IDF, information retrieval
Topics: Comparative (2 or more geographical areas), Global, English, BCE-4th Century, 15th-17th Century, 19th Century, attribution studies and stylometric analysis, text mining and analysis, Linguistics, Literary studies

The proposal focuses on keywords extraction; its aim is two-fold. Firstly, the paper provides an evaluation of the existing techniques, namely log-likelihood keyword analysis, Zeta as developed by Burrows and refined by Craig, as well as TF-IDF weighting. Secondly, the paper introduces a brand-new method of extracting meaningful keywords, which relies on a simple observation that ordered word frequencies provide enough information about particular words' potential keyness.

THE LUSOPHONE DIGITAL HUMANITIES AND WHAT THEY (WE) ARE DOING FROM THE SOUTH TEXTUAL CORPUS ANALYSIS AND FAIR PRINCIPLES TO TACKLE HEGEMONY.

Ricardo Medeiros Pimenta
ricardopimenta@ibict.br
Instituto Brasileiro de Informação em Ciência e Tecnologia

Priscila Ramos Carvalho
Universidade Federal do Rio de Janeiro

Josir Cardoso Gomes
Instituto Brasileiro de Informação em Ciência e Tecnologia

Keywords: Digital Humanities, FAIR, Google Scholar, South American, Iramuteq.
Topics: South America, Global, English, Contemporary, information retrieval and querying algorithms and methods, meta-criticism (reflections on digital humanities and humanities computing), Humanities computing, Library & information science

Introduction

The year 2018 witnessed a significant growth of research groups and laboratories dedicated to Digital Humanities in Brazil, however, without producing for this international community. Nowadays, the Digital Humanities are beginning to gain greater public interest in Brazil and other countries in South America.

In this perspective, our research would like to discuss what is produced in the Lusophone-speaking Digital Humanities in South America. The difficulty in diagnosing such production is evident because the global information regime has surrendered to the socio-technical and cultural monopoly mediated by Google, Amazon, Facebook, Apple, and Microsoft, major

technological players (Fiormonte & Sordi, 2019). In the case of the Portuguese-speaking world, it is noticeable the language barrier often puts the debate and dialogue-less in evidence. In addition, English literature as it is evidently English-speaking escapes from the larger question of problematization in the face of critical thinking, which is decolonization.

Methodology

The present work came from the research data of Gomes et al (2018) that retrieved Digital Humanities academic papers, thesis, and books written in the Portuguese language from Google Scholar. Despite Google Scholar's public access, it does not apparently provide consistent means that meet FAIR principles - Findable, Accessible, Interoperable, Reusable (Wilkinson, 2016), due to concentration and opacity of information retrieved, that it may be visible but not operable.

This empirical study analyzed 454 abstracts which composed the textual corpus through text mining techniques with IRaMuTeQ - Interface "R" for Multidimensional Analysis of Texts and Questionnaires (Marchand & Ratinaud, 2012).

Results

The result of the similitude analysis (graph theory) unveiled possible thematic convergences of the Portuguese language production in the Digital Humanities. The graph showed a central cluster represented by the term digital, which has a semantic attraction with the following terms: conceito, texto, meio, ferramenta, linguagem, and interação. In addition, this cluster has two subclusters identified by artigo and processo, as well it links other opposing clusters: novo, estudo, and pesquisa. On the top right, the cluster novo has two subclusters social and nao. In the first subcluster, the term social presented a possible connection between the highlighted terms: comunicacao, informacao, social, and rede. The second subcluster demonstrated a balance between the terms: internet, possivel and acesso. On the bottom left, the cluster estudo exposed the term analisar in evidence, and it links to the cluster pesquisa which is on the extremity of the graph. From this analysis, it can infer the term set reflects actions, products, secondary research objects, and methods, besides the problems and challenges of non-internet access in the South.

The word cloud allowed the quick visualization and identification of the main keywords of the textual corpus: digital, analisar, pesquisa, novo, comunicacao, pesquisa, tecnologia and nao. Furthermore, this result reinforced the perception of the similitude analysis.

Conclusions

The usual bibliometric retrieval based on Web of Science (WOS) and Scopus databases does not show the plethora of academic papers produced on Global South. From a decolonizing perspective, this study shows that scraping Google Scholar data could bring a broader result if you want to analyse portuguese scientific production. In addition, the use of Zenodo allowed the research result to have a visibility to the public outside Brazil allowing that the South American production could integrate Lusophone Digital Humanities in the global context, thus representing an important and necessary technopolitical action for researchers from that language community.

References

1.Gomes, J.; Castro, F.; Pimenta, R. (2018). Google Scholar como fonte de medição da produção científica lusófona. LATmetrics - Anais do I Congresso de Altmetria e Ciência Aberta na América Latina. Dataset accessible: .

2.Fiormente, D. & Sordi, P. (2019). Humanidades Digitales del Sur y GAFAM. Para uma geoppol[ítica del conocimento digital. Liinc em Revista, Rio de Janeiro, v.15, n.1, pp.108-130.

3.Marchand, P. & Ratinaud, P. (2012). L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française. Actes des les Journées Internationales d'Analyse Statistique des Données Textuelles, pp.687-699.

4. Wilkinson, M. D., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3. DOI: 10.1038/sdata.2016.18

Experiments in Analysis of Industry—Academia Collaboration and Research Trends

Daniil Skorinkin
National Research University Higher School of Economics, Moscow, Russia

Irina Busurkina
National Research University Higher School of Economics, St. Petersburg, Russia

Oleg Serikov
oserikov@hse.ru
National Research University Higher School of Economics, Moscow, Russia

Keywords: academic research analysis, social network analysis, natural language processing
Topics: Comparative (2 or more geographical areas), Global, English, 20th Century, Contemporary, natural language processing, spatial & spatio-temporal analysis, modeling and visualization, History of science, Sociology

The algorithm of research submission and publication is strict and highly formal. This produces lots of metadata accompanying research papers. The analysis of such datasets is not a mainstream research field nowadays. Yet such an analysis could help us better understand certain large-scale patterns of contemporary academic publishing, and maybe even the dynamics of science.The paper presents experiments in mix-method analysis of academic research papers and respective metadata. We investigate a corpus of abstracts and metadata for 31, 000+ papers from the arXiv.org e-prints archive. Our experiments reveal the properties of interinstitutional research and the collaboration of the industry and academy. We also do visualize and interpret the temporal properties of research trends.

Le Lotus and the women's voices of Egypt

Amélie Alrifaee
aalri099@uottawa.ca
Université d'Ottawa, Canada

Keywords: Women authors, feminism, gender studies
Topics: Africa, English, 20th Century, cultural analytics, digital archiving, Feminist studies, Literary studies

Between 1892 and 1919, over 30 journals and periodicals written by women for women were publish in Egypt. These publications paved the way to the Middle East's most successful feminist movement which took place in Egypt after the revolution of 1919. The period during which these journals and periodicals were published and the feminist movement they fueled is known as the Women's Awakening. For my poster presentation, I want to highlight the relation between the women's press of Egypt and the political and literary landscapes prior to the 1919 revolution by analyzing the texts of the Lotus.

A Survey of Data and Encodings in Word Clouds

Muyang Shi
shim@carleton.edu
Carleton College, United States of America

Danielle Albers Szafir
University of Colorado Boulder, United States of America

Eric Carlson Alexander
Carleton College, United States of America

Keywords: word clouds, visualization
Topics: Global, English, Contemporary, media archaeology, meta-criticism (reflections on digital humanities and humanities computing), Cognitive sciences and psychology, Computer science

Word clouds are an increasingly popular means of presenting statistical summaries of document collections, appearing frequently in digital humanities literature, newspaper articles, and social media. Despite their ubiquity and intuitive appeal, our ability to read such visualizations accurately is not yet fully understood. Past work has shown that readers perform poorly at certain tasks with word clouds, and that perceptual biases can affect their interpretation. To better understand the potential impacts of these biases, we present a survey of word cloud usage. Drawing from a corpus of literature from the fields of digital humanities, data visualization, and journalism, we record what data encodings are most commonly used (e.g., font size, position, etc.), what data is being presented, and what tasks are meant to be supported. We offer design recommendations given the most common tasks and biases, and point to future work to answer standing questions.

Network Analysis Finds Shifts in the History of Modern Architecture

Gregory Yauney
gjy24@cornell.edu
Cornell University

David Mimno
Cornell University

Keywords: networks, architecture, history
Topics: Europe, English, North America, 20th Century, network analysis and graphs theory and application, Art history, Computer science

Architectural historical texts implicitly create networks of relations between the objects, people, and texts they examine. Histories of European modern architecture examine a consistent core of these entities. From each of four English-language histories of European modern architecture (by Le Corbusier, Reyner Banham, Anthony Vidler, and Jean-Louis Cohen), we construct a graph of the objects, people, and texts mentioned in the same context. Different modeling resolutions can be used; we add undirected, unweighted edges between all nodes that occur on the same printed pages. We construct our graphs at the page-level because the data comes from book indices. In actor-network theory, which is popular in many fields including architectural history, as in our work, both humans and cultural artifacts can be network constituents. Networks have been used in other digital art history, for example in social analysis and mapping projects. To our knowledge, ours is the first work on architectural history that explicitly constructs networks from texts and quantifies aspects of them.

These four graphs provide overlapping but distinct views of modern architecture spaced throughout the last century. Network analysis, in the form of eigenvector centrality and triangle counting, allows us to make observations about both the form and content of these histories. This process allows comparisons between how the different histories relate groups of objects, and we find that some of the texts reexamine objects in multiple contexts much more frequently. Our results compare and contrast how the works construct historical importance, which is determined through close readings of the histories in question and supplemented by other histories of modern architecture. We quantify the importance of different people, objects, and texts in each of the networks, providing an empirical comparison between stated notions of historical importance and importance derived from the structures of the texts. Network analysis provides additional evidence to augment rather than supplant traditional ideas of historical importance.

We find that certain architects that the histories portray as important, like Le Corbusier and Walter Gropius, are in fact the most central in the four graphs. In the later histories, we find that modern architects are more often mentioned with the work they influenced rather than with the works to which they understood themselves to be responding. For example, Le Corbusier wrote about his response to Greek architecture, while the later histories focus on his relation to buildings other architects designed during his early career. Methodologies for comparing networks derived from multiple texts are used to investigate architectural history in this case study, but they are also useful for any situation with multiple views of relationships between similar entities. We expect the methodology to afford larger historiographical studies in other areas of architecture and history.

The data is publicly available: https://github.com/gyauney/modern-architectural-history-indices

An Open Data Approach to Revealing Indigenous Texts in Large-Scale Digital Repositories: A Case-Study of Locating Pages of Māori Text in the HathiTrust

David Bainbridge
davidb@waikato.ac.nz
University of Waikato, New Zealand

J Stephen Downie
jdownie@illinois.edu
University of Illinois, USA

Hemi Whaanga
University of Waikato, New Zealand

Keywords: Low-resource languages, Content discovery, Te Reo Māori
Topics: Global, English, Australia/Oceania, 19th Century, 20th Century, Contemporary, information retrieval and querying algorithms and methods, open access methods, First nations and indigenous studies, Library & information science

In this case study we report on our experiences in locating pages of Māori text in the HathiTrust Digital Library (HTDL). Using traditional biographic metadata, i.e., the language field, only 182 items were returned out of HTDL's 17.1 million volumes. Our Open Data approach is based on the freely available HathiTrust Extracted Features Dataset. We establish a collection of high frequency terms in Te Reo Māori, which we iteratively use as search terms to identify a group of candidate texts. We then apply NLP analysis to verify those texts that contain substantial amounts of the Māori language. Using this approach we were able to increase the number of volume returned to 598. This positive result suggests that scholars who want to analyse other low-resourced languages should be able to adopt our workflow to reveal otherwise hidden texts in their desired languages.

Making Ukiyo-e Easier to Discover A Recommender System for Digital Archives

Jiayun Wang
jiayunwong@hotmail.com
Graduate School of Information Science and Engineering, Ritsumeikan University, Japan

Biligsaikhan Batjargal
biligsaikhan@gmail.com
Kinugasa Research Organization, Ritsumeikan University, Japan

Akira Maeda
amaeda@is.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

Kyoji Kawagoe
kawagoe@is.ritsumei.ac.jp
College of Information Science and Engineering, Ritsumeikan University, Japan

Ryo Akama
rat03102@lt.ritsumei.ac.jp
College of Letters, Ritsumeikan University, Japan

Keywords: Digital Archive, Ukiyo-e, Recommender System
Topics: Asia, English, Contemporary, artificial intelligence and machine learning, information retrieval and querying algorithms and methods, Computer science, Library & information science

Ukiyo-e is a kind of woodblock print that has high artistic and research value. It is preserved by many digital archives (DAs), such as the Art Research Center Ukiyo-e Portal Database (ARC-UDB) of Ritsumeikan University. ARC-UDB is mainly built for the experts of humanities fields. In this research, to meet the potential needs of the expert users who will browse or explore ARC-UDB, we propose a recommender system. The proposed recommender system utilizes an existing link prediction model, which exploits the graph-like datasets of ARC-UDB as the input of the recommendation algorithm. We optimize the format of input of the link prediction model, to the format that is suitable for ARC-UDB datasets. From the results, we find that the proposed method is effective for the task. This recommender system could also be applied to other DAs that are with graph-like dataset structures.

Conceptual Modeling of Similarities and Duplication in Large-Scale Digital Libraries

Maggie Ryan
margaret.ryan@du.edu
University of Denver, United States of America

Lindsay Gypin
University of Denver, United States of America

Krystyna K. Matusiak
University of Denver, United States of America

Benjamin M. Schmidt
New York University, United States of America

Peter Organisciak

University of Denver, United States of America

Keywords: text-mining, text analysis, digital libraries
Topics: English, North America, Contemporary, digital libraries creation, management, and analysis, text mining and analysis, Library & information science

Large-scale digital libraries, such as the HathiTrust Digital Library (HTDL) and the Internet Archive have emerged consortially, collecting works from institutions around the world. This has led to unevenly biased duplication: some works recur many times in the collections, while others may only have one copy. The Massive Text Lab at the University of Denver is researching levels of 'sameness' and duplication of works within these digital libraries through massive-scale analysis. We will discuss applications to modern cataloging standards and provide an overview of the issue and intricacies of duplication, the solutions the project is pursuing, and the value that our work provides in framing material relationships for future humanities scholarship.

Feminist Revisions: An Emergent Model for DH in Libraries

Anne Cong-Huyen
annech@umich.edu
University of Michigan, United States of America

Miranda Marraccini
miramarr@umich.edu
University of Michigan, United States of America

Caitlin Pollock
pollockc@umich.edu
University of Michigan, United States of America

Keywords: service design, feminist praxis, libraries, critical pedagogy, care
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, sustainable procedures, systems, and methods, Education/ pedagogy, Library & information science

How can a library develop a collaborative digital scholarship service model and embed feminist critical pedagogy into our teaching, research, and support activities? As an alternative to these existing models for digital humanities support in university libraries, our poster outlines a feminist, anti-racist, and equity-centered approach that we used in the formation of a digital scholarship services pilot at the University of Michigan Library. To ground the poster in concrete examples, we focus on the three core programs that are developed and coordinated from within the DS Hub: (1) service model design, (2) digital pedagogy support, and (3) open data workshops and scholarship support. Through these three examples, we will demonstrate how these feminist and anti-racist principles manifest in practice around the values of consent, trust, and care.

Automatic Labeled Data Generation for Person Named Entity Disambiguation on the Ming Shilu

Richard Tzong-Han Tsai
thtsai@csie.ncu.edu.tw
Department of Computer Science and Information Engineering, National Central University, Taiwan
Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

Cheng-Han Wu
Department of Computer Science and Information Engineering, National Central University, Taiwan

Pi-Ling Pai
Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

I-Chun Fan
Institute of History and Philology, Academia Sinica, Taiwan

Keywords: Named Entity Disambiguation, Automatic Labeled Data Generation, BERT
Topics: Asia, English, 15th-17th Century, artificial intelligence and machine learning, natural language processing, Computer science, History

One important task of historical research in DH is to identify person names from history texts. This task can be divided into two subtasks: person named entity recognition (PNER) and person named entity disambiguation (PNED). PNED is to link each PNE mention to a specific person profile in the reference knowledge base. The main challenge of machine-learning-based PNED is the lack of annotated data. We design an automatic approach to labeling the training data. We choose the Ming Shilu as our target history texts. We use the Ming-Qing Archives Name Authority Database as our reference knowledge base, which contains 14,070 government officials living in Ming dynasty. Our BERT-based model reaches an accuracy of 90.1%, which proves that our approach can generate labeled data for the PNED task of very high quality on Chinese history texts. For the general situation (including trivial instances), the accuracy is even higher (~98%).

MAZI means together: An open-source "minimal computing" local network infrastructure used for cultural event organisation, fieldwork research and community-based curation

Mariana Ziku
ziku@mail.com
History of Art Laboratory, Department of Fine Arts and Art Sciences, School of Fine Arts, University of Ioannina, Greece

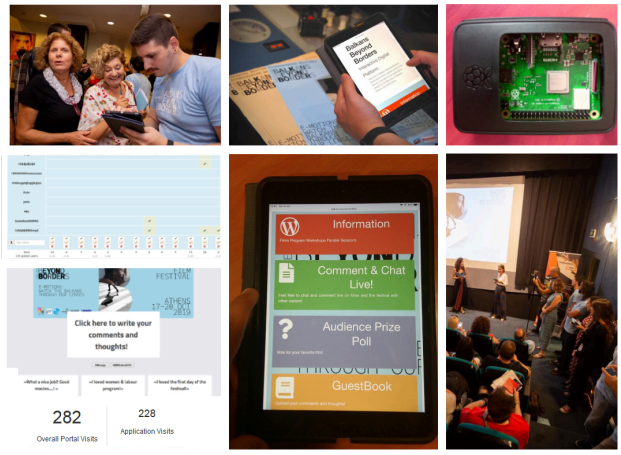
Elli Leventaki
History of Art Laboratory, Department of Fine Arts and Art Sciences, School of Fine Arts, University of Ioannina, Greece

Keywords: minimal computing, community networks, local wireless network, digital curation
Topics: Comparative (2 or more geographical areas), Europe, English, Contemporary, digital ecologies and digital communities creation management and analysis, physical & minimal computing, Computer science, Cultural studies

The emergence of Community Networks (CNs) and DIT (Do-It-Together) minimal computing ecosystems has resulted in technological solutions that enhance community connectivity and digital inclusion. The case is made for the cultural uses of local network infrastructures that combine wireless technology, low-cost hardware, and free/libre/open source software (FLOSS) applications. Based on these features, the toolkit MAZI ("together" in Greek), a Horizon 2020 project initiated by NITlab, University of Thessaly, Greece, has been deployed for creating pop-up local wi-fi zones independent from the internet, that enable digital interactions of communities within a low physical proximity coverage range. In this context, the focus is to explore the cultural-technological intersectionality of local community networks and its affordances as useful infrastructures for enhanced cultural event planning, humanities research and curatorial practices.

The objectives of ICT-enabled local networking as a research, curatorial and communication tool within the scope of humanities, digital scholarship and the GLAM sector, can be directed to foster new participatory curatorial forms, the digital documentation of transient -off the internet- community knowledge sharing and inspiring alternative experiences of the locality and commonality.

The three applied cases presented here are examples of physical proximity community networking platforms that have adjusted and utilised the open-source applications of the MAZI toolkit (NextCloud, Etherpad, LimeSurvey and Wordpress) in different cultural settings, mounting the toolkit on a low-sized hardware with minimal computing capabilities (Raspberry Pi):



i. Media exchange, audience communication and voting in a cross-cultural Balkan event

The toolkit was used for enhancing audience engagement during a multifaceted short film festival in the Balkan region. A platform was developed for sharing images, comments, chatting with local community members and enabling a voting system where individual preferences could be expressed. MAZI operated as a hands-on tool for building a participatory digital infrastructure for cultural event organisation, particularly suitable for outdoors or out of internet range areas.

ii. Collaborative commenting and anonymous participation in community-based fieldwork research

During community-based research, we employed an array of participatory techniques that elicit multimodal qualitative data: collective drawing, collaborative creative writing, reflective blogging, storytelling. Protecting anonymity, ensuring privacy and ethical data recording was a critical part of the research project. We employed the build-in anonymity features of MAZI that allow users to connect and share without registering their identities (by default). In this setting, MAZI beared the potential to transform a group of people into a convivial, spontaneous and creative research community, producing critical and ready to analyze empirical data.

iii. Digital exhibition hosting and community-based curation with added content

Digital exhibitions in HTML format built from scratch, hosted in a variety of open-source local networking infrastructures (Piratebox, Librarybox, MAZIzone). The exhibitions could be accessed only locally, on-site, traveling along the venues. The audience could explore the exhibition by connecting to the local network (no internet access) through their personal devices, with the option to upload their own content, collectively curating and contextualising the exhibition. The use case of MAZI here was utilising wireless local networks as exploratory digital curatorial tools.

Bibliography:

Antoniadis, P. (2016). Local networks for local interactions: Four reasons whyand a way forward. First Monday, 21(12). doi:10.5210/fm.v21i12.7123

Davis, G. Gaved, M. (2017). Seeking togetherness: moving toward acomparative evaluation framework in an interdisciplinary DIY networkingproject. In: Proceedings of the 8th International Conference on Communitiesand Technologies (C&T '17), Association for Computing Machinery, pp.96-99.

Dragona, D. (2015). "From Community Networks to Off-the-Cloud Toolkits", in I. Theona and D. Charitos (eds.) Hybrid City Conference III: Data to the People. Athens: URIAC

GO:DH Minimal Computing working group. "Minimal Computing" Retrieved from https://go-dh.github.io/mincomp/thoughts/2016/10/02/minimal-definitions

Gurstein, M. 2007. What is community informatics (and why does it matter)? Polimetrica, Milan, Italy.

NewsEye: A digital investigator for historical newspapers

Antoine Doucet
University of La Rochelle

Martin Gasteiner
University of Vienna

Mark Granroth-Wilding
University of Helsinki

Max Kaiser
National Library of Austria

Minna Kaukonen
University of Helsinki

Roger Labahn
University of Rostock

Jean-Philippe Moreux
National Library of France

Guenter Muehlberger
University of Innsbruck

Eva Pfanzelter
University of Innsbruck

Marie-Eve Therenty
University Paul Valéry Montpellier

Hannu Toivonen
University of Helsinki

Mikko Tolonen
mikko.tolonen@helsinki.fi
University of Helsinki

Keywords: Historical newspapers, quantitative analysis
Topics: Global, Europe, English, 19th Century, 20th Century, artificial intelligence and machine learning, natural language processing, History, Library & information science

The NewsEye H2020 project, running from May 2018 until April 2021, is an interdisciplinary undertaking that involves 3 European national libraries, 4 humanities and social science research groups and 4 computer science research groups. The core concept of NewsEye is a seamlessly integrated armory of tools and methods that will improve the users' capability to access, analyze and use the content in the digital libraries of historical newspapers.

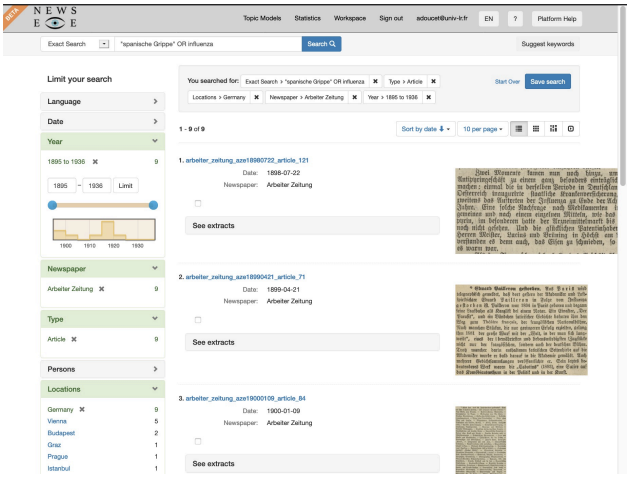


Figure 1: Beta version of the NewsEye demonstrator (June 2020)

Specifically, in the context of historical newspaper written in German, Finnish, Swedish and French, with a focus on the period 1850-1950, the project aims to develop a toolbox consisting of two main layers, as well as novel research results on several topics and in several fields of digital humanities, based on documents in different languages, so as to demonstrate the potential extent of its usefulness as a catalyst for the seamless development of novel research.

In details, the first layer of the NewsEye toolbox focuses on tools to improve and enrich historical newspapers, with improved text recognition and article segmentation, followed by semantic enrichment through the recognition and linking of named entities, stance detection, as well as novelty detection. A

language-independent set of higher quality data results from this step, already allowing an enriched experience and access to the newspaper collections. The second layer of the toolbox provides ways to benefit from this enriched dataset, through dynamic text analysis tools interacting with respect to user activities: contextualized topic modeling, viewpoint and comparative analysis, etc. In addition, an innovative personal research assistant is able to design strategies (plans) for finding something interesting and to revise them on the fly when needed. It consists of an investigator (dynamically finding and suggesting novel ideas), a reporter (summarizing the grounds for all suggestions) and an explainer (allowing the user to understand the suggestions by herself, and to return to the original data to confirm or infirm them).

Within the project, several digital humanities case studies are led, with the aim to guide the development of adequate tools, and so as to demonstrate their potential for the development of novel research in digital humanities. In the case studies, groups of humanities scholars carry out investigations for representative research issues, such as "gender", "migration", "nationalism and revolutions", and "media". Since there is plenty of existing qualitative research on these topics, the project strives towards making an impact in the fields of historical research and digital humanities by combining knowledge from qualitative analyses with new findings in big data analyses provided by the new tools in this project.

Figure 2: Short description of the case study on return migration

It is essential to understand that the NewsEye research topics and datasets are showcases, and that the seamless inclusion of additional research question is a key ambition. With this in mind, all the tools developed are language-independent, so as to be able to seamlessly integrate further datasets. In fact, NewsEye is both open to further research cases to be studied using its tools and to the integration of additional datasets and tools through the status of associated partner.

Acknowledgements:

This work has been supported by the European Union Horizon 2020 research and innovation programme under grant 770299 (NewsEye).

Further information:

NewsEye project Website in 4 languages:
https:// www.newseye.eu/ [de] [en] [fi] [fr]
NewsEye platform: https://platform.newseye.eu/
NewsEye twitter account:
https://twitter.com/NewsEyeEU
NewsEye project publications and data sets:
https://zenodo.org/ communities/newseye/
NewsEye project software and source code:
https://github.com/NewsEye/
NewsEye podcasts:
https://www.univie.ac.at/newseye/

NewsEye videos: https://www.youtube.com/channel/UCwEqOk8JRfbJeBYV-ZkPctA/playlists

Text Analysis of 19th and 20th Century Etiquette Books Using R

Laura Biesiadecki
tuf84653@temple.edu
Temple University, United States of America

Keywords: Text Analysis, Literature
Topics: English, North America, 19th Century, 20th Century, text mining and analysis, Feminist studies, Literacy, composition, and creative writing

In the interest of contributing to a session on "carrefours/intersections," I propose a poster reviewing the progress of a text analysis project: during the 2019-2020 academic year, I will be working with more than 250 etiquette books and conduct guides for men, women, and children, published in America between 1848 and 1920, the goal being to identify the ideal American woman as she was written and revised between the landmark convention at Seneca Falls and the ratification of the 19th Amendment.

The visual representation of this project will highlight the intersection between digital resources and long-form academic projects in the humanities, emphasizing the accessibility of these digital methods to graduate students as well as the potential benefit of using code to analyze large swaths of extratextual data.

A comparative study of sentiment and topics in migration related tweets

Suzanne Alayne Moody
alayne.moody@mail.mcgill.ca
McGill University, Canada

Sunyam Bagga
McGill University, Canada

Keywords: Sentiment Analysis, Topic Modeling, Social Media, Twitter, Migration
Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, social media analysis and methods, text mining and analysis, Computer science, Media studies

The International Organization for Migration (IOM) recently reported that the world has more migrants than ever before "both numerically and proportionally" and that the number of environmental migrants alone could reach 1 billion by 2050 (2017, p. 2; 2014). The IOM has also argued that understanding how people think and feel about migration is essential to the development of policy that supports the safe passage and successful integration of migrants into their new communities (2017). The United States receives more migrants per year than any other country (United Nations, 2016). It is also a country where migration is a prominent topic in social, political and media discourse. This poster reports the findings of an exploratory study whose principal research questions are: 1) What do Twitter users in the United States talk about in their migrant- and migration-related tweets, 2) how do authors feel when they tweet about these subjects, and 3) how do their sentiments and topics of interest compare to Twitter users residing outside the United States? To approach these questions, we apply sentiment analysis, topic modeling and time series analysis to 111,785 English language tweets containing the term "migrant" or "migration" collected from the Twitter API between January 21 and March 4, 2019. Our quantitative and computational methodology relied on descriptive statistics, the Python programming language, and associated libraries, such as the Natural Language Toolkit (NLTK) for text processing, Pandas for data manipulation, VADER for sentiment analysis, Gensim for topic modeling and Seaborn for data visualization. The visualized data shows that authors in the United States focused more on politics whereas authors in other countries focused more on humanitarian issues. Tweets from U.S. authors were more negative than those authored by residents of other countries, except when the topics of the tweets involved children. A time series plot revealed three sentiment spikes, one positive and two negative, over the course of the collection period. The negative spikes were partially explained by looking at word frequencies, visualized as word clouds, alongside news headlines for the dates in question. The positive spike was more difficult to explain because of the limitations of sentiment analysis and the negative nature of news reporting. Our findings cannot be generalized because the Twitter

Search API does not generate a representative sample (González-Bailón, 2014). Future work could involve an improved sampling method that would permit the use of inferential statistical methods. The addition of multiple languages to the tweet dataset or social network data to the analysis may yield interesting new findings.

References

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. Social Networks, 38, 16-27. https://doi.org/10.1016/j.socnet.2014.01.004

International Organization for Migration. (2014). IOM Outlook on Migration, Environment and Climate Change. https://environmentalmigration.iom.int/iom-outlook- migration-environment-and-climate-change-1

International Organization for Migration. (2017). World Migration Report 2018. https://www.iom.int/wmr/world-migration-report-2018

United Nations, Department of Economic and Social Affairs, Population Division (2016). International Migration Report 2015: Highlights (ST/ESA/SER.A/375). https://www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/MigrationReport2015_Highlights.pdf

Studying Geographical Patterns Across John Milton's Genres

Randa El-Khatib

elkhatib.randa@gmail.com
University of Victoria, Canada

Keywords: Milton, Python, geography, genre
Topics: Global, Europe, English, 15th-17th Century, spatial & spatio-temporal analysis, modeling and visualization, text mining and analysis, Geography and geo-humanities, Literary studies

This research project employs Python's Natural Language Toolkit (NLTK) to study John Milton's references to old and new geography throughout his career and across his wide-ranging genres. This project investigates whether the blend of old and new geography is genre specific, and traces location-based patterns to study whether Milton's geographic focus is also genre specific

RavenSpace Books Unbound

Darcy Cullen

cullen@ubcpress.ca
UBC Press, Canada

Beth Fuget

bfuget@uw.edu
University of Washington Press, United States of America

Amber Ridington

ridington@ubcpress.ca
UBC Press, Canada

Keywords: digital publishing, Indigenous scholarship, collaborative methodologies, decolonizing education
Topics: Global, English, North America, 19th Century, 20th Century, Contemporary, digital publishing projects, systems, and methods, public humanities collaborations and methods, First nations and indigenous studies, History

This poster addresses tensions that arise in publishing Indigenous studies scholarship in a digital environment. It takes RavenSpace as its case study, a new publishing platform for media-rich, networked, interactive books in Indigenous studies where communities and scholars can work together to share and create knowledge. Developed by UBC Press along with the University of Washington Press and other partners, and funded by the Andrew W. Mellon Foundation, the platform respects Indigenous protocols for accessing and using cultural heritage and traditional knowledge while meeting the standards of peer-reviewed academic publishing. It enables authors to present their own work and to recontextualize materials already online in accord with Indigenous protocols for attribution, control, care, and sharing. It offers multiple paths through projects to reach different audiences.

Indigenous communities and scholars increasingly shape their research to serve the needs and priorities of the communities involved, including their cultural preservation and revitalization efforts, educational initiatives, and capacity building goals, among others. As the M?ori scholar and educator Linda Tuhiwai Smith writes, "Indigenous research seeks to ... find the new/

old solutions that restore Indigenous being in the world. The impacts are part of an intergenerational long game of decolonization, societal transformation, healing and reconciliation, and the recovery of a world where all is well."[1] Of relevance to the public digital humanities and open data movements that some participate in, Indigenous scholars and community experts want to present their research in innovative ways that engage and are tailored to the various scholarly, community, and public audiences they aim to reach. In light of the historical misappropriation of Indigenous cultural heritage, they are also concerned with developing protocols for accessing and protecting that heritage in the digital domain.[2]

There is an ongoing tension between Indigenous concerns and the open data movement within public digital humanities, since digital "repatriation" has often resulted in Indigenous materials being dumped online, decontextualized and open to the world. Traditional publishing models also fail to meet new needs. Neither the old authorial hierarchies nor the limitations of the printed book do justice to the research currently being carried out by Indigenous communities and scholars. How can publishers develop new modes of publishing that respond to contemporary scholarship and learning in Indigenous studies, are accountable to the Indigenous peoples involved in the projects, and respect Indigenous protocols for the control, use, and sharing of their cultural knowledge and materials?[3]

This poster focuses on the technical and social infrastructure developed for this purpose in RavenSpace. It demonstrates the tools and methods designed to address these questions in a digital space, including adaptable approaches to establishing protocols for the respectful use of cultural content and processes that support collaborative authorship. It explains how the academic publishers involved in RavenSpace have expanded the peer review process to include both academic and community review. It also point people to the first RavenSpace publication, As I Remember It, by Sliammon Elder Elsie Paul with Davis McKenzie, Paige Raibmon, and Harmony Johnson, published in 2019.[4]

[1] Linda Tuhwei Smith, "The Art of the Impossible - Defining and Measuring Indigenous Research?," in Dissident Knowledge in Higher Education, ed. Marc Spooner and James McNinch (Regina: University of Regina Press, 2018), 38. See also: Margaret Kovach, Indigenous Methodologies: Characteristics, Conversations, and Contexts.(2009) University of Toronto Press.; Shawn Wilson, Research Is Ceremony. 2008, Fernwood Publishing.

[2] See, for example, Deirdre Brown and George Nicholas, "Protecting Indigenous Cultural Property in the Age of Digital Democracy: Institutional and Communal Responses to Canadian First Nations and Maori Heritage Concerns," Journal of Material Culture 17, no. 3 (2012): 307-24; Kimberly Christen, "Does Information Really Want to Be Free? Indigenous Knowledge Systems and the Question of Openness," International Journal of Communication 6 (2012): 2870-93, http://ijoc.org/index.php/ijoc/article/view/1618; Mark Turin, "Ownership, Control, Access & Possession in Open Access Humanities Publishing." ScholarLed 10-21-2019 [accessed October 22, 2019] https://blog.scholarled.org/ownership-control-access-possession-in-oa-humanities-publishing/; David Gaertner and Karyn Recollet, "Indigenous Digital Pedagogy." Digital Pedagogy in the Humanities: Concepts, Models, and Experiments. Modern Language Association (2018) [accessed October 22, 2019] https://digitalpedagogy.mla.hcommons.org/keywords/indigenous/; Robin Boast and Jim Enote, "Virtual Repatriation: It Is Neither Virtual nor Repatriation." Heritage in the Context of Globalization, Springer Briefs in Archaeology (2013) 8: 103-113; Jane Anderson and Kimberley Christen. "'Chuck a Copyright on it': Dilemmas of Digital Return and the Possibilities for Traditional Knowledge Licenses and Labels." Museum Anthropology Review (2013) 7:105-126; Kate Hennessy, "Cultural Heritage on the Web: Applied Digital Visual Anthropology and Local Cultural Property Rights Discourse." International Journal of Cultural Property (2012) 19:345-369; and Amber Ridington and Kate Hennessy, "Building Indigenous Agency Through Web-Based Exhibition: Dane-Wajich - Dane-zaa Stories and Songs: Dreamers and the Land." Proceedings of Museums and the Web 2008 [accessed October 22, 2019] https://www.museumsandtheweb.com/mw2008/papers/ridington/ridington.html.

[3] Darcy Cullen and Alan Bell, "The Social Text and Networked Knowledge: New Modes of Scholarly Book Publishing in Indigenous Studies." Journal of Scholarly Publishing 49(2), 193-212.

[4] [accessed October 22, 2019] http://publications.ravenspacepublishing.org/as-i-remember-it

"Webs of Violence in Beowulf"

Alexandra Bolintineanu
alexandrabolintineanu@gmail.com
University of Toronto, Canada

Keywords: Beowulf, Old English, violence, narrative
Topics: Europe, English, 5th-14th Century, network analysis and graphs theory and application, rhetorical analysis, Literary studies

Editions of the Old English poem Beowulf open with family trees, networks of birth and alliance. This project uses the network graph tool Cytoscape to visualize and analyze the network of who kills whom in the poem. The project examines the ethical and rhetorical decision-making that underlies a translation of Beowulf's poetics of violence into the visual language and mathematical aspects of network graphs. The shape of the resulting network graph, with its connected sub-regions, mirrors the narrative structure of the poem. Not only is violence pervasive; it structures every single narrative, whether foreground or episode. Violence is the procedural logic of Beowulf. By representing the poem as a network graph, the project stages a "new media encounter zone," in Alan Liu's formulation, between Beowulf's inset oral storytelling, Beowulf's print editions' paratext, and the affordances of the bioinformatics visualization platform Cytoscape.

Lightning Talks

Opening up Open Data Strategies & success stories

Vanessa Hannesschläger
vanessa.hannesschlaeger@oeaw.ac.at
Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Austria

Tanja Wissik
Austrian Centre for Digital Humanities, Austrian Academy of Sciences, Austria

Keywords: Open Data, Open Source, Hackathon, Community involvement
Topics: Global, Europe, English, Contemporary, open access methods, public humanities collaborations and methods, Humanities computing

Not only in the context of Digital Humanities but also in other research areas the Open Data movement is gaining momentum. For this reason, the Austrian Centre for Digital Humanities of the Austrian Academy of Sciences (ACDH) started an experiment at the beginning of 2019: For the first time ever, we published the calls for participation in three virtual hackathons funded by the Austrian manifestations of DARIAH and CLARIN, CLARIAH-AT. These hackathons focused on Open Data sets that are publicly available online, and the tasks to perform on these data involved the creation of Open Source code. Each of the hackathons had a special theme and was co-timed with an event that involved an aspect of Openness. These events also inspired the choice of the respective data sets.

Usually hackathons take place on site, participants are given tasks to be solved within a given timeframe in a fixed location. This requires the programmers to be flexible and available and to have access to travel funding. A virtual hackathon on the other hand offers people all over the globe the possibility to participate and contribute without having to travel. Therefore, our approach enabled a much larger community to participate in the event on the one hand, thus also promoting the benefits of Open Data on the other.

The first hackathon was carried out in cooperation with the European Lexicographical Infrastructure (ELEXIS) and focused on lexicographical data (the Digital Dictionary of Tunis Arabic). The task was to develop a creative mode of processing it, e.g. by enriching it, visualizing it, doing statistical analysis, or integrating it with other resources (e.g. LOD). The submissions were to be handed in by the end of the first ELEXIS observer event taking place in February 2019.

For the second hackathon, the ACDH cooperated with the City of Vienna and chose a set from the city's Open Government Data platform to be processed. The task was to be completed on the Open Data Day Vienna 2019 (28 February 2019). The aim was to develop a creative mode of processing cartographical data showing damage to buildings in Vienna during World War II.

The third and final hackathon of the series offered a task to be completed on the International Open Data Day 2019 (2 March 2019). For this final highlight, a choice of two Open Data sets was offered to the participants. The first data set to be worked on in this task was a collection of XML/TEI transcriptions of early German travel guides. The second data set consisted of German historical newspapers and was provided in cooperation with the Europeana newspaper project.

The best contributions were determined by an international board of judges and received cash prizes. The criteria for judgement were creativity and innovation, accessibility, reusability and reproducibility, as well as elegance. In our presentation, we will share the lessons learned and show how Open Science was the necessary precondition for this project, as well as what inspired its ultimate success. We will make our case by giving insights into the lessons learned in 2019 and sharing how we improved the concept for the second round of the hackathon series in 2020.

Bibliography

ACDH-CH. ACDH Virtual Hackathon Series. 2019. https://www.oeaw.ac.at/acdh/detail/event/acdh-virtual-hackathon-series/

ACDH-CH. ACDH-CH Open Data Virtual Hackathon - Round Two. 2020. https://www.oeaw.ac.at/acdh/detail/event/acdh-ch-open-data-virtual-hackathon-round-two/

Vanessa Hannesschläger. The ACDH virtual hackathon series: Open Data for Open Source solutions. DARIAH Open Blog, 23. May 2019. https://dariahopen.hypotheses.org/571

PRESERVATION OF OSUN-OSOGBO CULTURAL HERITAGE IN NIGERIA: A DIGITIZATION DISCOURSE

ISRAEL ABAYOMI SAIBU
alerosaibu@gmail.com
ANCHOR UNIVERSITY, LAGOS NIGERIA.

AYOMIDE JOSEPH SAIBU
ayomidesaibu2001@gmail.com
LAGOS STATE UNIVERSITY OJO, LAGOS NIGERIA

Keywords: Preservation, Digitization, Cultural Heritage, Osun-Osogbo, Nigeria
Topics: Africa, English, 19th Century, Contemporary, digital archiving, digitization (2D & 3D), Art history, History

The Osun-Osogbo cultural heritage has become one of the most striking cultural identities to have emerged in Nigeria. The uniqueness and importance of Osun-Osogbo has led to its recognition by the United Nations Educational Scientific and Cultural Organization (UNESCO) as a global cultural heritage. Thus, it is imperative that this cultural heritage be digitized for the preservation of its essence for posterity and global visibility.

The Rabbinic Social Network

Michael Satlow
michael_satlow@brown.edu
Brown University, United States of America

Michael Sperling
New York University, United States of America

Keywords: historical networks, gephi, pattern recognition
Topics: Asia, English, BCE-4th Century, 5th-14th Century, network analysis and graphs theory and application, History, Theology and religious studies

This project attempts to apply the techniques of social network analysis (SNA) and visualization to the representations of rabbinic interactions in the Babylonian Talmud, a sprawling text written in Hebrew and Aramaic and probably redacted in Babylonia (modern day Iraq) in the sixth century CE. Our goals are (1) to develop a workflow and methodology allowing us to visualize and analyze the interactions between rabbis as represented in the Babylonian Talmud; (2) to see if we could learn something new about the relationship between rabbis as represented in the Talmud and/or the process of its redaction; and (3) to present a public-facing interface allowing scholars to interact directly with our visualization.

Many of the research questions that drive this project go back more than a century. Pioneering work in Jewish studies (especially Albeck (1969); Margolioth (1987)) has attempted to detail the relationships between some of these rabbis. This work remains valuable, although it sometimes uses outdated methodological assumptions. Some of the relationships, for example, are reconstructed on the basis of stories about rabbis that most scholars today would understand as late, fictional creations. So too, scholars have long tried to understand the process by which the Babylonian Talmud was redacted (for summary of the scholarship on this, see Rubenstein 2013). Historians have also tried to understand the rabbis as a "network", although without applying the quantitative tools now available (Hezser (1997); Lapin (2012)).

In this part of the project, we focused our attention on citation chains. Rabbis frequently say things in the name of other rabbis (e.g., "Rabbi X said in the name of Rabbi Y who said in the name of Rabbi Z" – these chains usually consist of two or three names but can go up to nine!). By focusing on simply the names in these chains (and not the content of what they reported), our work intersects with that of Zhitomirsky-Geffet and Prebor (2018) and Josh Waxman (2019). At the same time, both our workflow and the kinds of questions that we were asking of the network as a whole make it distinct.

The first step in our workflow was to identify each instance of a citation chain in the Hebrew/Aramaic text. In order to do this, we compiled a list of the names of all rabbis mentioned in ancient rabbinic literature (along with any aliases that they had) and assigned each a unique numeric identifier. The list was created through both automated and manual processes. Then, we created and ran a pattern matching program on a digital version of the "standard" printed edition (Vilna) of the Babylonian Talmud text to identify instances of rabbinic names and citation chains. Once identified, the program split the citations into "source" and "target" rabbis so we could identify who was citing whom. The results of the automated process were highly accurate as we verified through manual review of a statistically significant sample.

The program identified 5,245 citation instances. When grouped into unique interactions (e.g., Rabbi X may cite Rabbi Y twenty times, but we counted that as one unique interaction), we were left with 630 rabbis

(our nodes) and 1217 unique interactions (our edges). We loaded our node and edges files into Gephi (Gephi) and UCINET. A visualization can be seen in Figure 1, which (using a Force Atlas 2 layout) groups the more connected rabbis toward the center.

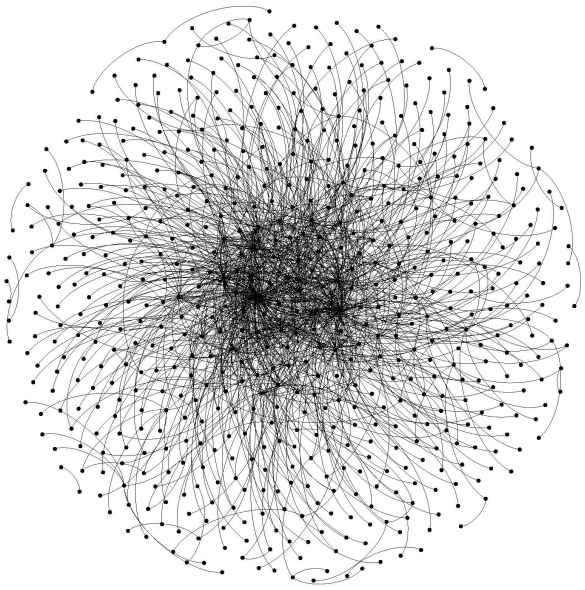


Figure 1: Graph of All Rabbis in Babylonian Talmud Who Appear in Citation Chains

We have two major research findings. First, and less surprisingly, when separated into Modularity Classes through an unsupervised algorithm, the rabbis relatively cleanly separated into groups that clustered around rabbinic figures who themselves had many connections, which looks like a "school" structure (see Figure 2). Previous research has led us to expect this. Second, and more surprisingly, the rabbis at the centers of each of those circles were themselves densely and directly connected to each other. These ten or fifteen rabbis, over four centuries and two geographical locales, served as the backbone for the rabbinic network. It is still unclear to us whether these connections represent real social interactions or can better be explained as the result of later editing and redactional decisions and conventions.

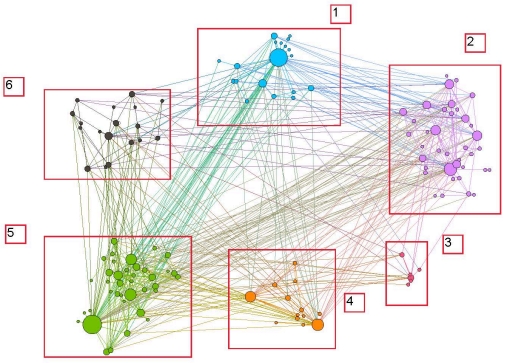


Figure 2: Rabbis in Citation Chains in Babylonian Talmud in Modularity Classes

There are problems inherent in this data. The rabbis are themselves sometimes unsure about an attribution (and explicitly argue about it). Since rabbis sometimes shared names or nicknames, it is sometimes impossible to match with certainty a name with a distinct individual; in such cases we assigned the shared name to the more prominent rabbi. Moreover, we are just using one, easily available text. Manuscripts sometimes record these attributions differently. We feel that given the macro approach we took to this network, these problems become less significant. Nevertheless, they need to be better taken into account in future, more fine-grained analyses.

By the time this is published, we should have our data and code freely available on Github. It may take us longer to develop a public-facing interface, perhaps along the lines of "The Six Degrees of Francis Bacon." We will also extend our approach to other interactions in the Babylonian Talmud (e.g., when a rabbi asks a question of another rabbi); to other rabbinic texts from this period; and to other manuscript versions of these texts.

Bibliography

Albeck, Ch. 1969. Introduction to the Talmud Babli and Yerushalmi. Tel Aviv: Dvir (in Hebrew).

Gephi: https://gephi.org/

Hezser, C. 1997. The Social Structure of the Rabbinic Movement in Roman Palestine. Tübingen: Mohr Siebeck.

Lapin, H. 2012. Rabbis as Romans: The Rabbinic Movement in Palestine, 100-400 CE. New York: Oxford University Press.

Margolioth, M. 1987 (rpt). Encyclopedia of Talmudic and Geonic Literature. Tel-Aviv: Y. Orenstein (in Hebrew).

Rubenstein, J. 2013. "Translator's Introduction," in David Weiss Halivni, The Formation of the Babylonian Talmud (New York: Oxford University Press), xvii-xxx.

Six Degrees of Francis Bacon: http://www.sixdegreesoffrancisbacon.com/?ids=10000473&min_confidence=60&type=network

UCINET: https://sites.google.com/site/ucinetsoftware/home

Waxman, J. "Mi vaMi: A Graph Databse of the Babylonian Talmud." http://www.mivami.org/

Waxman, J. 2019. "A Graph Database of Scholastic Relationships in the Babylonian Talmud": https://www.narcis.nl/dataset/RecordID/hdl%3A10411%2FK8HHQZ

Zhitomirsky-Geffet, M. and Prebor, G. 2019. "SageBook: Toward a Cross-Generational Social Network for the Jewish Sages' Prosopography." Digital Scholarship in the Humanities 34.3: 676-695 (https://doi.org/10.1093/llc/fqy065)

Interactive Spatial Positioning of Historical Photographs in Virtual Reality Results of a User Study

Florian Niebling
florian.niebling@uni-wuerzburg.de
University of Würzburg, Germany

Anna Riedmann
University of Würzburg, Germany

Keywords: Virtual Reality, gamification, geo-spatial visualization

Topics: Europe, English, 19th Century, 20th Century, Contemporary, spatial & spatio-temporal analysis, modeling and visualization, virtual and augmented reality creation, systems, and analysis, Art history, Computer science

University-Industry Partnerships in the Humanities: View from the partner and academic perspective

Lynne Siemens
siemensl@uvic.ca
University of Victoria

Keywords: collaboration, academic teams, project management

Topics: English, North America, Contemporary, project design, organization, management, Humanities computing

While university-industry partnerships are common in the sciences, they are more rare on the humanities side. This provides an opportunity to explore one such collaboration based in the humanities. In this case, exploring open scholarship, the partnership involves libraries, academic-adjacent organizations and academics. Through a series of interviews, these parties express that they find the experience positive and see both benefits and challenges, albeit from different perspectives. For example, libraries and academic-adjacent organizations focus on learning while the academics are interested in moving research to production. The challenges include a focus on cultural differences and the partners' ability to navigate these. In terms of measures of success and desired outcomes, they are both in agreement that these measures and outcomes are soft in nature, though focused on influencing government policy on social scholarship. The partnership members will continue to invest time, resources, and intellectual capacity to the endeavor.

Computation and Rhetorical Invention: Finding Things To Say With word2vec

Mark Wolff
wolff.mark.b@gmail.com
Hartwick College, United States of America

Keywords: rhetoric, invention, word2vec, machine learning
Topics: Europe, English, North America, Contemporary, artificial intelligence and machine learning, electronic literature production and analysis, Literacy, composition, and creative writing, Literary studies

In his recent book Friending the Past, Alan Liu laments the waning of a rhetorical regime that until recently had held sway in literary studies as a means of making sense of the past. Instead of using what Liu names as rhetoric-representation-interpretation to convey an understanding of history, we are now stuck in an ambiguous regime of communication-information-media where it is not clear how we reach an understanding of anything (2-3). The shift from rhetoric-representation-interpretation to communication-information-media is not unique in the history of literary studies, however. It follows another shift that occurred over 100 years ago in how rhetoric was deployed. Gérard Genette observed in 1966 that literary studies had not always emphasized representations. Before the end of the nineteenth century, literary studies revolved around the art of writing. Texts were not objects to interpret but models to imitate: students demonstrated their understanding of literature by mastering elocution and reproducing figures of style in the works they read. With the institution of literary history as a nationalist project at the end of the nineteenth century, academic reading approached texts as objects to be explained according to prescribed methods for documenting how literature represented a national identity. This new way of studying literature stressed disposition, or the arrangement of ideas in the service of ideology.

The methods of literary history would eventually be used by literary scholars in the twentieth century to turn narratives about literature away from nationalism toward other priorities, most notably poststructuralism and the critique of cultural hegemonies.

Recent developments in information technology have further challenged paradigms for reading literature. Digital tools for text analysis allow for the study of large corpora using quantitative methods. As Ted Underwood, Andrew Piper and others have shown, large-scale computational text analysis has called into question fundamental concepts in literary history such as periodization, nationality, and genre. Using computational methods can enable us to develop models for literary studies, but these models are not limited to interpretation. Computational techniques such as topic modeling and word vector spaces can facilitate investigations into the possibilities for literary creation. Technology has the potential for exploring invention, or the finding of ideas to express through language given a context that can be parameterized.

If, in literary studies, an emphasis on elocution or style served the perpetuation of social hierarchies, and if an emphasis on disposition or argumentation challenged these hierarchies by promoting forms of knowledge and ideologies, a new rhetorical emphasis is needed to respond to the ontological condition of the communication-information-media era. We are surrounded by data with no clear way to make sense of it, and we need to explore inventional methods of finding things to say within this state of being. Digital environments today constitute in part the material context for suasive activity, and as Thomas Rickert argues, contemporary rhetoric must attend to how humans and the world are in this context (xv). The affordances of networked access to texts and computational processing contribute to a rhetorical ambience that grants a degree of agency to the environment in what is said about the world, which includes literature.

I will consider two examples of how tools for computational literary studies lend themselves to inventional practices. The first is ReRites, a year-long project by David (Jhave) Johnston who used a neural net trained on various corpora to produce poetry ("Why A.I.?", 172). The raw text of the poetry was generated by computation but Johnston edited the output. In terms that emphasize the materiality of computation, Johnston describes his role as "carving the text." Neither the computer nor Johnston writes these poems in the sense we usually give to writing: they emerge from the world in which a machine and a human find themselves. After performing complex analyses on very large corpora, the machine produces something the human takes to find something to say with language.

The second is SonGenApp, a web application I developed that enables a user to select verses from a large corpus of sonnets to assemble a new poem. From all the verses in the corpus a word embedding is

modeled with word2vec, and from the model the user selects verses semantically with an analogy based on a pair of words. The user can modify a selected verse as long as it follows the rules of scansion and rhyme for sonnets. With the application attending to formal constraints, the task of the user is to find verses that are meaningful in some way at the moment of using the application. The user can always read the source texts for selected verses and base the construction of the generated poem on a knowledge of literary themes and history. But this prior knowledge is not necessary. The user can encounter verses in the corpus by changing the analogy as if it were a knob on a black box.

The quantity of digital texts at our disposal opens possibilities for discovery in rhetorical invention. Stephen Ramsay has described a "hermeneutics of screwing around" where browsing resources leads serendipitously to the pleasure of finding things one had not anticipated. Computation has the potential to afford the same discovery in finding things to express through writing.

Works Cited

Genette, Gérard. "Rhétorique et Enseignement." Figures II, Editions du Seuil, 1969, pp. 23-42.

Johnston, David (Jhave). ReRites: Machine Learning Poetry Edited by a Human. http://glia.ca/rerites/. Accessed 8 Oct. 2019.

--. "Why A.I.?" ReRites?: Human + A.I. Poetry?; Raw Output?: A.I. Trained on Custom Poetry Corpus?; Responses?: 8 Essays about Poetry and A.I., edited by Stephanie Strickland. Anteism Books, 2019, pp. 171-76.

Liu, Alan. Friending the Past?: The Sense of History in the Digital Age. University of Chicago Press, 2018.

Piper, Andrew. Enumerations: Data and Literary Study. U of Chicago P, 2018.

Ramsay, Stephen. "The Hermeneutics of Screwing Around; or What You Do with a Million Books." Pastplay: Teaching and Learning History with Technology, edited by Kevin Kee, University of Michigan Press, 2014, pp. 111-20.

Rickert, Thomas. Ambient Rhetoric: The Attunements of Rhetorical Being. U of Pittsburgh P, 2013.

Underwood, Ted. Distant Horizons: Digital Evidence and Literary Change. U of Chicago P, 2019.

Wolff, Mark. mbwolff/SonGenApp. 2019. GitHub, https://github.com/mbwolff/SonGenApp.

Coding Literary Ecologies

Elizabeth Anne Swanstrom
swanstro@gmail.com
University of Utah

Keywords: environmental humanities, digital humanities, literary studies, natural language processing, media ecologies
Topics: English, North America, 19th Century, 20th Century, Contemporary, eco-criticism and environmental analysis, natural language processing, Literary studies, Environmental, ocean, and waterway studies

Literary scholarship tends to treat the natural world as the stage upon which narrative unfolds, rather than the essential pre-condition for existence. "Coding Literary Ecologies" brings the environmental features of literary texts into detailed relief in a playful yet strategic manner. It is comprised of four web-based applications, each accompanied by an essay that situates it within literary history and contextualizes it within contemporary discourse about the digital and environmental humanities.

The Montias Case: an experiment with data reconciliation and provenance between research and cultural heritage institutions

Leon van Wissen
l.vanwissen@uva.nl
University of Amsterdam

Chiara Latronico
c.latronico@uva.nl
University of Amsterdam

Sandra van Ginhoven
Getty Research Institute

Veruska Zamborlini
University of Amsterdam

Keywords: Reconciliation and Disambiguation of Linked Open Data, Provenance, Named Entity Recognition in Archival Sources, Cooperation public and research institutions, Dutch Golden Age
Topics: Europe, English, North America, 15th-17th Century, 18th Century, digital archiving, linked (open) data, Art history, History

This paper discusses the complex process of reconciliation of data coming from research and public cultural heritage institutions with their own selection criteria that have shaped the provenance of their collections. We demonstrate how the Golden Agents digital humanities research infrastructure [1] can play an intersecting role as intermediary data provider between these distributed collections in the reconciliation, disambiguation and deduplication of data by taking their provenance into account. To this end we analyse an art historical case we call 'The Montias Case', with data from three different sources: the Getty Provenance Index, [2] the Frick Collection, [3] and the notarial acts of the Amsterdam City Archives. [4]

In the 1980s, John Michael Montias (1928-2005) began to compile a database containing records of ownership of works of art of the Dutch Golden Age. To this end he also selected records from the Amsterdam City Archives using inventories dated 1597-1681. In 1985 and 1987, Montias was invited to the Getty Research Institute to automate his work and incorporate it into the Getty Provenance Index. When his collaboration with the Getty concluded around 1990, Montias continued his project and donated his work to the Frick Art Reference Library. Both datasets were enriched by others. The Getty Provenance Index and the Frick Collection have now two partially overlapping datasets with a shared provenance that are in need of deduplication/disambiguation to fully exploit them for art-historical research.

The Getty Research Institute and the Golden Agents consortium set up an experiment in 2019 to reconcile the Montias data with their provenance and enrichments to their original source: the notarial acts. This reconciliation via the Golden Agents infrastructure is necessary because the Amsterdam City Archives due to their public mission only index person and geographical names, but not the art objects that Montias listed that we need as researchers.

We first link the inventories from the Getty Provenance Index and the Frick Collection to their corresponding notarial deeds. Additionally, we deduplicate the names indexed in these data sources and identify people utilizing the Lenticular Lenses II tool (Idrissou et al., 2018, 2019) in development within the Golden Agents project. Then we link the individual items indexed in the Montias database to the transcription of the historical archival source. The Web Annotation Model is used as a data model for the text and serves both modelling and presentational purposes. This allows us to link and to validate the exact piece of text indicated by coordinates on the scan where a certain object is listed and to present it for instance in a IIIF-environment. The thesauri AAT and TGN, [5] and ICONCLASS [6] help in describing these items in a standardized way. Finally, we demonstrate how, in line with the goals of the Golden Agents and the remodel of the Getty Provenance Index as Linked Open Data projects, [7] the results of the Montias Case can be relevant for data reconciliation and for modeling the provenance of distributed heterogeneous cultural heritage collections for (re-)use in digital humanities research projects in general.

[1] https://www.goldenagents.org

[2] https://www.getty.edu/research/tools/provenance/search.html

[3] https://research.frick.org/montias/

[4] http://archief.amsterdam/archief/5075/

[5] https://www.getty.edu/research/tools/vocabularies/

[6] http://www.iconclass.org/

[7] https://www.getty.edu/research/tools/provenance/provenance_remodel/index.html

References

Bibliography

Al Idrissou, Veruska Zamborlini, Chiara Latronico, Frank van Harmelen, Charles van den Heuvel. Amsterdamers from the Golden Age to the Information Age via Lenticular Lenses. DHBenelux 2018, 6-8 June. Amsterdam, The Netherlands.Al Idrissou, Frank van Harmelen, Veruska Carretta Zamborlini, Chiara Latronico. Contextual Entity Disambiguation in Domains with Weak Identity Criteria: Disambiguating Golden Age Amsterdamers. K-CAP 2019, ACM, New York, USA (forthcoming).

Web

Art & Architecture Thesaurus (AAT) and Getty Thesaurus of Geographic Names (TGN), https://www.getty.edu/research/tools/vocabularies/ .Frick Collection, https://research.frick.org/montias/ .Getty Provenance Index, https://www.getty.edu/research/tools/provenance/search.html .Getty Provenance Index Remodel Project, https://www.getty.edu/research/tools/provenance/provenance_remodel/index.html .Golden Agents, https://www.goldenagents.org .Iconclass, http://www.iconclass.org .Notarial Acts of the Amsterdam City Archives, http://archief.amsterdam/archief/5075/ .

"Digitizing Rochester's Religions Piloting a Community–University Partnership in the Digital Humanities."

Daniel James Gorman Jr.
dgormanj@ur.rochester.edu
University of Rochester, United States of America

Keywords: Religion, Urban, Public, History, Archive
Topics: English, North America, 19th Century, 20th Century, Contemporary, digital archiving, public humanities collaborations and methods, History, Theology and religious studies

Launched by Dr. Margarita Guillory at the University of Rochester in fall 2016, Digitizing Rochester's Religions documents the evolution of religion in Western New York after the Second Great Awakening (1800-1850) ended. Western New York and the city of Rochester were renowned for revivals and new religious movements during the Second Great Awakening, so that the region became known as the "Burned-over District." However, Western New York's religious history after 1850 has not received equal scholarly attention. Dr. Guillory and the graduate and undergraduate students who worked on DRR sought to fill this gap. (I served as the lead graduate student researcher.) The team wrote essays about past and present religious sites, visited religious sites and archives, and digitized sources from community archives. We sought to collaborate with local religious communities, so that DRR would build a meaningful relationship between the University and surrounding neighborhoods.

Bruce Lincoln's definition of religion, namely that a religion consists of a "discourse," a "set of practices," a "community," and "an institution," guides DRR. [Source: Bruce Lincoln, Holy Terrors: Thinking about Religion after September 11 (Chicago: University of Chicago Press, 2003), 5-7; see 5 for "discourse," 6 for "set of practices" and "community," and 7 for the full quote of "an institution."] In its attention to religious spaces and its collaborative approach to scholarship, DRR draws inspiration from Dr. Courtney Bender's Sacred Gotham, which tasked students with mapping religious spaces in New York City. It also builds on Dr. David H. Day's 2003 web project "Encountering Old Faiths in New Places: Mapping Religious Diversity in the Rochester, New York Area," based at Monroe Community College (https://web.archive.org/web/20071102133941/http://www.monroecc.edu/depts/sociology/pluralism/overview.htm). "Encountering Old Faiths" featured students' ethnographic observations of current religious sites. DRR, by contrast, profiles past as well as present religious sites, so that it is more historical than anthropological in its orientation. Overall, DRR contributes to a growing field of public-facing projects about lived religion in U.S. cities; notable examples include Boston's Hidden Sacred Spaces (http://www.hiddensacredspaces.org/) and the American Religious Sounds Project (https://religioussounds.osu.edu/).

When Dr. Guillory moved to Boston University in 2018, I took over DRR and completed it as a pilot project, documenting religious communities in Rochester's southwester quadrant, instead of the whole city as originally planned. The website (http://digrocreligions.org/) provides a template for historians, religionists, and students to pursue this work on a larger scale. Taken together, the essays featured on DRR detail how, beginning in the 1960s, the loss of Rochester's industrial base exacerbated racial and economic segregation. Religious organizations in the economically distressed southwestern neighborhoods filled the gap left by the withdrawal of tax dollars and government services. By launching job programs, soup kitchens, and clinics, religious groups in southwest Rochester tried to meet the physical and material as well as spiritual needs of residents.

Had COVID-19 not required the cancellation of DH2020, my lightning talk would have reviewed DRR's public and digital history aspects and provided a tour of the website. I would have discussed the importance of developing reciprocal relationships with community partners (in our case, religious congregations), although we were not fully successful. After I took over the project, the priority became finishing it, so that I only used publicly available documents instead of archival sources to finish several essays. As for digital humanities technology, I would have discussed the workflow of scanning, formatting, cataloguing, and compressing 80 gigabytes' worth of primary sources for the website. Finally, I would have presented DRR as an example of successful project-based learning, since the students who worked on DRR gained hands-on archival, ethnographic, and digital humanities experience.

La Enciclopedia de Banrepcultural: Una enciclopedia digital de patrimonio cultural colombiano, que nace de la colaboración de los visitantes de las Bibliotecas del Banco de la República

Irene Tobón Restrepo
itobonre@banrep.gov.co
Banco de la República de Colombia, Colombia

Keywords: Colombia, cultural heritage, crowdsourcing
Topics: South America, English, Spanish, 19th Century, 20th Century, Contemporary, crowdsourcing, digital biography, personography, and prosopography, Art history, History

La Enciclopedia de banrepcultural.org es una publicación digital de conocimiento sobre el patrimonio cultural colombiano y en español. Es de libre acceso, libre uso y gran parte de su contenido nace de una creación colaborativa a través de campañas de crowdsourcing.

Inició en el año 2000 con respuestas a las consultas académicas de los visitantes a la Biblioteca Luis Ángel Arango y se creó un listado de cerca de 700 artículos que respondían a estas preguntas. En 2017 se decidió publicar estos contenidos en un gestor MediaWiki y se crearon nuevas convocatorias para la generación de información.

En esta nueva fase queremos aprender a integrar La Enciclopedia con Wikipedia y Wikidata, así como ampliar los proyectos colaborativos para traducir los artículos a otros idiomas y a las lenguas indígenas del país. Nos preocupa la poca presencia de conocimiento sobre Colombia y Suramérica que se encuentra disponible en internet.

«La planète numérique» d'un peuple autochtone transnational Une analyse des liens hypertextes des sites web amazighs

Abdelaziz BLILID
a.blilid@gmail.com
University of Tours, France

Keywords: peuples autochtones - culture amazighe - activisme culturel - communauté imaginée
Topics: Africa, French, Contemporary, cultural analytics, linked (open) data, African and African American Studies, First nations and indigenous studies

This research processing on the Internet using the Aboriginal Nations in the North America: les Berbères ou les Amazighs. Ce peuple a Investi Internet Internet dès les années 1990, a déclaré que le texte était représenté dans la toile numérique afin de permettre la transmission de la culture entre eux et d'associations et de faire en sorte que leurs revendications politiques (Almasude, 1994). Ils ont créé des sites Web pour codifier et transmettre leur patrimoine culturel. Ainsi, le Web leur a offert un moyen singulier de transmission: en transposant dans l'univers numérique, l'identité culturelle s'est redéfinie. This is this subject of this study at the people on the amazigh for cultural protection and activism is one one jou.

Quantitative Intertextuality: Analyzing the Markers of Information Reuse

Walter Scheirer
walter.scheirer@nd.edu
University of Notre Dame, United States of America

Christopher Forstall
cforstall@mta.ca
Mount Allison University, Canada

Keywords: Intertextuality, Text Analysis, Cultural Analytics, Computer Vision, Memes
Topics: English, North America, Contemporary, artificial intelligence and machine learning, text mining and analysis, Cultural studies, Literary studies

A remarkable amount of information crosses our eyes and ears each day, yet we adeptly identify what is familiar with seemingly no effort at all. In many cases, what we see or hear has been shaped by recognizable prior sources. Such instances of intertextuality reveal a wealth of data about authorship, influence, and style, making them attractive targets for automatic identification. This lightning talk introduces quantitative intertextuality [Forstall-and-Scheirer-2019], a new approach for the algorithmic study of information reuse in text, sound and images. Using a variety of tools drawn from machine learning, natural language processing, and computer vision, we will describe how to trace patterns of reuse across diverse sources for scholarly work and practical applications.

DIRECT - Digital Intertextual Resonances in Early Chinese Texts

Gian Duri Rominger
gianr@princeton.edu
Princeton University, United States of America

John O'Leary
jo10@princeton.edu
Princeton University, United States of America

Nick Budak
nbudak@princeton.edu
Princeton University, United States of America

Keywords: china, phonology, text, tool
Topics: Asia, English, BCE-4th Century, concordancing and indexing, natural language processing, Asian studies, Linguistics

In this talk, the creators will provide an introduction and demonstration of DIRECT (Digital Intertextual Resonances in Early Chinese Texts), a new digital tool designed to simplify phonological analysis of early Chinese texts. DIRECT enables users to quickly and easily identify homophonous and near-homophonous passages in ancient Chinese texts that may be obscured by graphic variation, uncovering hidden resonances and patterns.

Bringing Newcomers into the Fold: Faculty Development through Values-Driven DH Pedagogy

Kristen Mapes
kmapes@msu.edu
Michigan State University, United States of America

Ellen Moll
mollelle@msu.edu
Michigan State University, United States of America

Andy Boyles Petersen
andyjp@msu.edu
Michigan State University, United States of America

Keywords: faculty development, curriculum development, training
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, open access methods, Education/ pedagogy

A pervasive goal of DH is to bring new and curious faculty into the fold in a sustained way. At Michigan State University, we have developed a highly successful DH Pedagogy Learning Community to help new-to-DH instructors effectively integrate DH into their courses. Key to faculty engagement was our focus on how DH values such as collaboration, community, inclusion, public humanities, and experimentation aligned with their own priorities in teaching. We are now launching an OER that includes the curriculum for this values-driven learning community that can be used by either individual instructors or anyone wishing to lead their own discipline-agnostic learning community for new-to-DH faculty.

Análisis del coro como personaje en la dramaturgia grecolatina y española incluidas en DraCor

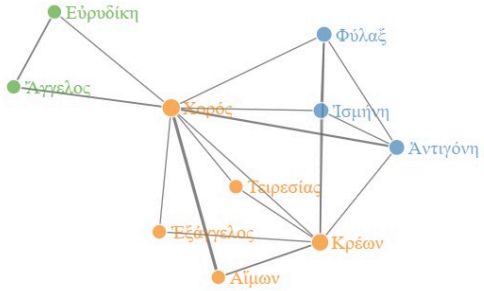
Teresa Santa Maria
teresa.santamaria@unir.net
Universidad Internacional de La Rioja, Spain

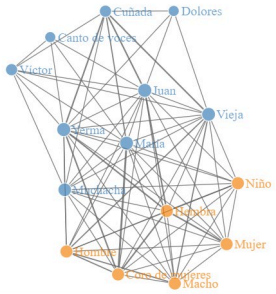
Monika Dabrowska
monika.dabrowska@unir.net
Universidad Internacional de La Rioja, Spain

Keywords: DraCor, Coro, Teatro, Grafo, Literatura comparada
Topics: Europe, Spanish, BCE-4th Century, 19th Century, 20th Century, network analysis and graphs theory and application, semantic analysis, Literary studies, Performance Studies: Dance, Theatre

Drama Corpora Project (DraCor) ha ido conformando un repositorio donde se encuentran, en estos momentos, once corpus teatrales diferentes, que pertenecen a periodos, géneros dramáticos y lugares geográficos diversos. Entre estos corpus encontramos veinticinco obras de la Edad de Plata española (finales del XIX-mediados del siglo XX) que conforman el "Spanish Drama Corpus" y que han sido importadas de la Biblioteca Electrónica Textual del Teatro Español de 1868-1936 (BETTE).

Además de los textos dramáticos en XML-TEI que se consultan en DraCor, podemos crear las redes sociales con grafos que conforman los personajes de cada una de dichas piezas dramáticas. Uno de los personajes comunes en muchas de estas piezas y que resulta fundamental en el desarrollo de la trama y para dar corporeidad a los pensamientos de otros personajes lo constituye el Coro.





What Is Formalisation? And Why Do We Need to Talk About It In DH?

Stefan Heßbrüggen-Walter
shessbru@hse.ru
National Research University Higher School of Economics

Keywords: formalisation, markup, dh software, philosophy
Topics: Global, English, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), Philosophy

The digital humanities as a discipline are centered on "formalisation". We acknowledge the fact that the machines we use employ formal languages, but we still lack a clear idea of what follows from this insight for the self-conceptualisation of the discipline. My presentation clarifies the concept of formalisation and spells out some of its practical consequences.

Building Online Communities as a Platform for Collaborative Learning and Cross-Cultural Self- Expression; Saudi Female Students' BlogHajj Behind the Scenes as an Example

Hadeer Aboelnagah
habouelnagah@psu.edu.sa
Prince Sultan University, Saudi Arabia

Keywords: Blogging, Collaborative Creativity, Sharing Knowledge, Open Access
Topics: Asia, Global, English, Contemporary, curricular and pedagogical development and analysis, public humanities collaborations and methods, Cultural studies, Education/ pedagogy

Digital Humanities as an emerging field provides endless opportunities to paradigm shifts in the educational experience in higher education, it opens wide doors for interdisciplinary collaborative students' projects that can be the seed of larger-scale national and international. Blogging is a pedagogical activity that is used to enhance collaborative learning and social responsibility (Yu-Chun Kuo 2017). Hajj Behind the Scenes is a blog by Saudi Female Students. The study explores blogging as an activity and its possible utilization to build cross-disciplinary online communities and to enhance societal involvement in the field of Digital Humanities. The said blog is used as an example to study its effect on the students and their abilities to self-expression, autonomous learning, and collaborative creativity.

Annotating spatial entities in Romanian Novels

Ioana Galleron
ioana.galleron@sorbonne-nouvelle.fr
Université Sorbonne-Nouvelle, Paris, France

Roxana Patras
Universitatea "Al. I. Cuza", Iasi, Romania

Camelia Gradinaru
Universitatea "Al. I. Cuza", Iasi, Romania

Keywords: spatial entities, Hajdouk novels, manual annotation
Topics: Europe, English, 19th Century, spatial & spatio-temporal analysis, modeling and visualization, text encoding and markup language creation, deployment, and analysis, Literary studies, Central/Eastern European Studies

This paper is based on HAIRO, a Franco-Romanian project for creating a library of Romanian Hajdouk novels in an XML/TEI format (see https://proiectulbrancusihairo.wordpress.com/home-1/). Hajdouks were outlaws living in the woods, that fascinated the public in the second half of the 19th century and at the beginning of the 20th century, both for their cruelty and their sense of justice. Between 1840 and 1920, they appear in almost 12% of the Romanian novels, with at least 40 titles specifically dedicated to this picturesque character.

Our main concern is "the place-making mediated by the text", and more precisely the creation of a Hajdouk space; in a rural Romania, structured by clear and stable relationships between spaces, their nomadic way of life constitutes a disrupting force, and we are looking at if and how this reflects in the novels. Much along the lines of (Hay and Butterworth 2019), our work focuses less on the "indexical relationship to the physical world", and more on the ways in which the texts create their own spatiality.

In the first part, we discuss the adaptation of Pustejovsky's ISO metamodel (2014, 2019) to operate what we call a "basic annotation" of our set of novels. Faced with the specificities of our texts, we have defined not two, but seven types of spatial entities: toponyms, places, paths, zones, vehicles, topical spaces and potential spaces. The two last categories are the most salient difference between our annotation schema and the previous existing ones, and we advocate their interest in literary contexts, where "the other world" or "in his bosom" are frequently mentioned, to quote but two examples from a very rich list.

We further characterize the spaces as "absolute" or "relative". For this "basic annotation", we have

renounced to define other types of relations, such as orientation, movement or metrics.

The annotation exercise took place in two phases. In a first, exploratory round, we have worked on XML files, and implemented our schema as a feature structure in TEI. In a second round, we have configured a BRAT server and started by measuring the inter-annotator agreement on a set of 10 samples of about 1000 words (see results in Galleron et al., forthcoming). In a third phase, currently under development, we proceed to the actual annotation of texts, using a place names dictionary to pre-annotate. Another path currently explored is that of the syntactic tagging of phrase constituents: since a large part of our space entities appear to assume a function of circumstantial complement of place, they could be spotted with a specialized dependencies tagger. However, the first experiences in this respect are quite disappointing, and all the more so they have been conducted on samples in French - results will probably be worse on Romanian samples, since Romanian is a language less equipped with NLP tools. Please note that usual NER systems (Stanford, Spacy library, etc.) do not work, or give very poor results, on Romanian texts. For all these reasons, manual annotation still appears as the best way to go, in spite of being extremely time consuming.

To date, the repartition of the annotations per type, as indicated in figure 1, confirms that looking at toponyms only, with a NER/ NEL approach, fails to capture a large part of the placemaking process in a novel. Also, two major categories of novels seem to appear with regards to the writing of the space, one constituted by the texts in which places and zones are in even proportions, the other gathering novels in which places are dominant, to the detriment of zones.

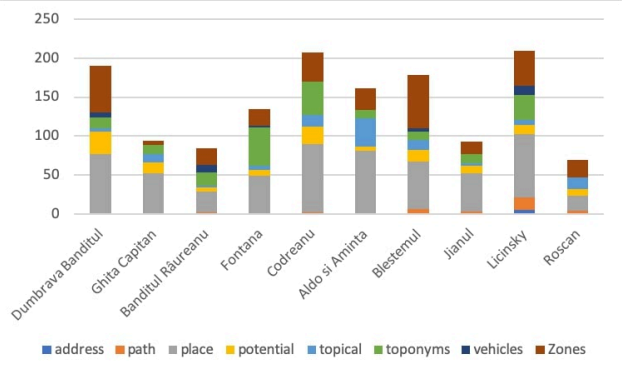


Figure 1. Annotations per type in a selection of novels

In addition, categories "paths" and "vehicles" seem to be discriminant between two other types of fiction. Indeed, while the number of annotations remains quite low in both cases, they allow to identify certain novels as outliers, with lots of spatial changes, as opposed to the major part of novels that appear finally more "static", and privileging scenes and summaries of the action. This is somewhat surprising, since we expected all our Hajdouk novels to pertain to the second category. We are currently trying to understand if the difference is motivated by the specific style of certain authors, the taste of an era, or it genuinely points towards a generic specificity within our corpus;

Bibliography

Bodenhamer, David J.; Corrigan, John; Harris, Trevor M., Deep Maps and Spatial Narratives, Bloomington & Indianapolis, Indiana University Press, 2015.

Bodenhamer, David J.; Corrigan, John; Harris, Trevor M., The Spatial Humanities: GIS and the Future of Humanities Scholarship, Bloomington: Indiana University Press, 2010, p. 109-123, muse.jhu.edu/book/26710.

Dicţionarul cronologic al romanului românesc, Bucureşti, Editura Academiei Române, 2003.

Galleron, Ioana, Patras, Roxana, Gradinaru, Camelia, Mélanie-Bécquet, Frédérique « À la recherche des haïdouks. L'annotation des entités spatiales dans un corpus de romans roumains du XIXe siècle », submitted to Revue des humanités numériques Humanistica, 2020.

Hay, Duncan, Butterworth, Alex (2019) "Spatial allusion, temporal recurrence and cognitive uncertainty: visualising chronotopic structure in a literary text", DH2019 Book of abstracts.

Pustejovsky, James, Lee, Kiyong, Bunt, Harry (2019) "The Semantics of ISO-Space", ISO committee draft for the revision of ISO (2014), https://let.uvt.nl/general/people/bunt/docs/PustejovskyLeeBunt_ISO-Space_ISA-15.pdf

Stop Words

Klara du Plessis

klaraduplessis@gmail.com
Concordia University, Canada

Keywords: stop words, graphs, soundwave visualization, poetry, Johanna Drucker
Topics: English, North America, Contemporary, image processing and analysis, meta-criticism (reflections on digital humanities and humanities computing), Literary studies

STOP WORDS is a set of five short poems based on five sets of commonly used stop words gleaned from topic modeling and computational literary studies: articles, pronouns, prepositions, conjugations of to be, and negation. The reading of each poem is recorded and each audio file is transformed using a variety of sound wave visualization softwares, resulting in five visual poems or images. These images function as interpretative graphical material, visualizing literature as simultaneously scientifically accurate, quantitative, and highly affective, qualitative, diagrams. At its broadest, Stop Words is also an intervention in the Humanities / Digital Humanities schismatic debate, digitally supplementing the traditional verbal dimension of poetry while celebrating an interpretative, subjective digitally generated product. Stop Words engages closely with Johanna Drucker's research on qualitative statistical representation.

Drucker, Johanna. Graphesis: Visual Forms of Knowledge Production. Cambridge, MA: Harvard UP, 2014.

Drucker, Johanna. "Humanities Approaches to Graphical Display." Digital Humanities Quarterly 5:1 (2011)

"Natural Language Toolkit." NLTK 3.4.5 Documentation. 2019.

Not TEI by Design: A Digital Scholarly Edition of Paradise Lost

Richard Cunningham
richard.cunningham@acadiau.ca
Acadia University, Canada

Keywords: Digital scholarly edition, John Milton, Paradise Lost, Encoding
Topics: Europe, English, North America, 15th-17th Century, Contemporary, digitization (2D & 3D), scholarly editing and editions development, analysis, and methods, Book and print history, Literary studies

The TEI continues to work toward becoming the default method for encoding and publishing digital scholarly editions. But TEI-xml has to be transformed, one way or another, in order to become readable by humans, either in print or online. To be read online, TEI encoded scholarly editions have to be transformed into the web's native language, html. Because "HTML and JavaScript have grown into . . . powerful technologies" (Ford and Thompson. "An Adventure with Client-Side XSLT to an Architecture for Building Bridges with Javascript") it makes sense, perhaps especially when presenting an already well-documented text, to eschew TEI and encode the text directly in html, using css and javascript (as, if needed php) to offer an affordance-rich digital scholarly edition. In this short presentation of _A Paradise Lost Portal_ such an edition, along with its encoding, will be offered.

Haikupedia

David F. Shultz
dshultz2@gmail.com
University of Toronto, Canada

Keywords: haiku, NLP, wikipedia, USE, deformance
Topics: Global, English, Contemporary, artificial intelligence and machine learning, natural language processing, Computer science, Literacy, composition, and creative writing

Haikupedia employs several "deformance machines" to transform Wikipedia into haiku. This radical contrast between form and content explores the use of hyper-condensed poetic form as a mode of "distant reading", and how genre expectations in haiku can be leveraged in machine translation to produce new works and modes of reading. One strictly formal deformance operation shapes sentences into a 5-7-5 haiku syllabic form ("An atom can be / ionized by removing one / of its electrons"); another loosens the 5-7-5 form (following Higginson) and imposes a phrase/fragment constraint (following Reichhold), implicitly enacting juxtaposition ("the couple / quickly fell in love / electrical signals"). In subsequent iterations, article text becomes "genetic material" in a genetic algorithm which uses convolutional neural networks and the Universal Sentence Encoder as parts of a complex fitness function.

Tiempo de destrucción: A stylometric examination of the posthumous publication of Luis Martín-Santos's unfinished work

Megan Cytron
mcytron@ucm.es
Universidad Complutense, Spain

Keywords: stylometry, literature, authorship, spanish
Topics: Europe, English, Spanish, 20th Century, attribution studies and stylometric analysis, text encoding and markup language creation, deployment, and analysis, Humanities computing, Literary studies

This study employs stylometric methods to disentangle questions of authorship and editorial intervention in a complex case study of the posthumously published novel, Tiempo de Destrucción. In 1964, a car accident cut short the life of the Spanish novelist Luis Martín-Santos, at a time when the literary community was anxiously awaiting the follow-up to his groundbreaking novel, Tiempo de Silencio (1961). A work in progress was found-according to Carlos Barral (of the publishing house Seix Barral): "Luis considered half of it to be almost definitive."

Littered with semi-legible marginalia and edits, the papers passed through many hands over the course of a decade. Using Rybicki, Hoover, and Kestemont's work on collaborative authorship and rolling data (2014) as a point of departure, I identify the areas where Martín-Santos's authorial signal is strongest and quantify the degree to which the editorial process and the text's incomplete, fragmentary nature interfere with Martín-Santos's authorial fingerprint.

Promoting Library Accessibility and Outreach with Special Collections in Virtual Reality

Julia DeWall Uhr
julia.uhr@colorado.edu
University of Colorado, United States of America

Keywords: VR, library, fairy-tales
Topics: English, North America, 18th Century, 19th Century, 20th Century, software development, systems, analysis and methods, virtual and augmented reality creation, systems, and analysis, Library & information science, Literary studies

This lightning talk will introduce the Special Collections of CU Fairytale Library in Virtual Reality (SCCUFL-VR), an open-source virtual library featuring books from the University of Colorado's collection of rare and historical fairy tale books, and will discuss the uses of SCCUFL-VR for pedagogy and outreach as well as SCCUFL's advantages over traditional digital collections.

Linking digitalized Buddhist scripts and Buddhist studies publications

Jen-Jou Hung
mctang@ntu.edu.tw
Dharma Drum Institute of Liberal Arts, Taiwan

Kuang-hua Chen
National Taiwan University, Taiwan

Muh-Chyun Tang
National Taiwan University, Taiwan

Keywords: Buddhist studies, CBETA, DLMBS, Social Network Analysis
Topics: Asia, English, Contemporary, digital research infrastructures development and analysis, network analysis and graphs theory and application, Library & information science, Theology and religious studies

In the study, we propose using data mining to automatically identify cited references to Buddhist texts available through CBETA (Chinese Buddhist Electronic Text Association) embedded in the full-text Buddhist studies publications collected by DLMBS (Digital Library and Museum of Buddhist Studies), which now includes

over 400,000 entries of Buddhist bibliographies in 45 languages, 65,000 of which are full-text. We believe that the cross-referencing between the Buddhist texts and related scholarly publications will provide great benefits for Buddhist scholarship. The users of the CBETA will have ready access to the relevant literature on the specific scripture they are studying. On the same token, the users of the DLMBS will be able to link directly to the exact scripture cited in the journal articles that they are reading. The aggregate co-citation data of Buddhist texts in the scholar publications also affords a rare opportunity to study the relationships among different Buddhist scriptures.

Creating Structured and Reusable Data for Tourism and Commerce Images of Edo Using IIIF Curation Platform to Extract Information from Historical Materials.

Chikahiko Suzuki

ch_suzuki@nii.ac.jp
ROIS-DS Center for Open Data in the Humanities, Japan National Institute of Informatics

Asanobu Kitamoto

ROIS-DS Center for Open Data in the Humanities, Japan National Institute of Informatics

Keywords: Edo (formerly Tokyo), pre-modern Japanese books, IIIF, curation, time machine project
Topics: Asia, English, 18th Century, 19th Century, data publishing projects, systems, and methods, digital research infrastructures development and analysis, Asian studies, Book and print history

Purpose

Our final research goal is to restore Edo's historical city conditions (Tokyo's former name before the late 19th century) using historical materials. In particular, we focus on non-textual materials. For textual materials, TEI is the de facto standard. For non-textual materials, we need to find new ways as just collecting images is insufficient. We need to organize images as data for analysis across historical materials.

Material and Method

The tool used for this work is the IIIF Curation Platform (ICP)[1]. "Curation" with ICP is a generic method that allows us to gather IIIF images or parts of images and add metadata to each. Curation allows images to be organized as structured and reusable data[2].

A large number of digitized pre-modern Japanese books are being released by multiple Japanese institutions[3]. We used ICP to organize 1255 images from eight guidebooks in the Edo period[4] then organized 2489 images from the Edo shopping guidebook[5].

Findings and Discussion

The current curated data confirmed the relationship between famous places and commerce. Presently, our finding reconfirms previous studies[6],[7]. The Asakusa area has many shops around Sensoji Temple today. However, there were many medicine dealers in one corner of Nihonbashi area unlike today (Figure 1-2).

By applying our method to Edo period publications, we not only structure useful information for our research, but also provide reusable data for other researches.

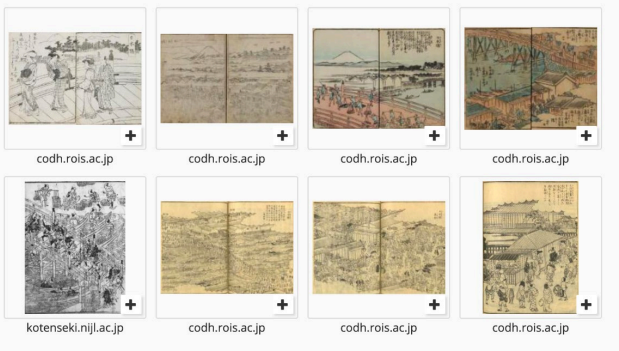


Figure 1: Nihonbashi area illustration curated from Edo guidebooks. Nihonbashi was one of the most famous places in the Edo period.

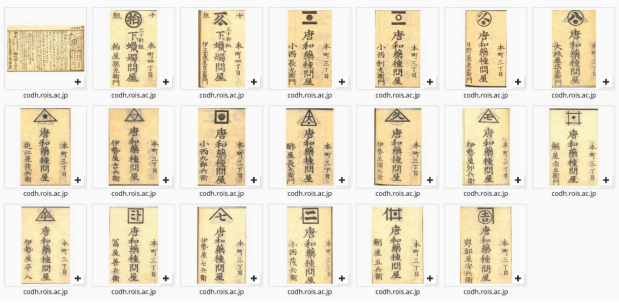


Figure 2: Advertisement of Nihonbashi merchant curated from the Edo shopping guidebook . There are many Chinese and Japanese medicine dealers 唐和薬種問屋 unlike today.

References

[1] IIF Curation Platform http://codh.rois.ac.jp/icp/index.html.en

[2] Suzuki, C. Kitamoto, A. (2019). "IIIF Curation Platform ni yoru Edo-Kaimono-Hitori-Annai no microcontents ka", Jinmonkon2019, pp.11-18

[3] Database of Pre-Modern Japanese Works, https://kotenseki.nijl.ac.jp/?ln=en

[4] Illustration of Edo Famous Places, http://www.ch-suzuki.com/icpedo/finder/?lang=en

Primary sources:

Asai, Ryoi. (1662). Edo-Meisho-Ki , Dataset of Pre-modern Japanese Works, doi: 10.20730/100000011

Kondo, Kiyoharu. (1663). Edo-Meisho-Hyakunin-Isshu, National Diet Library Digital Collection, doi: 10.11501/1183942

Burin, Shoufu. (1795). Ehon-Edo-Zakura, Dataset of Pre-modern Japanese Text, doi: 10.20730/200014989

Shinratei. (1795). Ehon-Edo-No-Mizu, Dataset of Pre-modern Japanese Text, doi: 10.20730/200013678

Saito, Choushu. (1834-1836). Edo-Meisho-Zue, Dataset of Pre-modern Japanese Text, doi: 10.20730/100249896

Utagawa, Hiroshige. (1850-1867) Ehon-Edo-Miyage, Dataset of Pre-modern Japanese Text, doi: 10.20730/200013687

[5] Curation of Merchant in Edo http://www.ch-suzuki.com/edoshop/finder/?lang=en

Primary source:

Nakagawa, Gorozaemon. (1824). Edo-Kaimono-Hitori-Annai, Dataset of Pre-modern Japanese Text, doi: 10.20730/100249503

[6] Ando, Y. (2005). Kankou Toshi Edo no Tanjo, Shinchosha.

[7] Yamamoto, M. (2005). Edo Kenbutsu to Tokyo Kanko, Rinsen Shoten

Deep Learning based Attribute Representation in Ancient Vase Paintings

Prathmesh Madhu
prathmesh.madhu@fau.de
Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg

Ronak Kosti
ronak.kosti@fau.de
Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg

Torsten Bendschus
torsten.bendschus@fau.de
Institute of Classical Archeology, Friedrich-Alexander-Universität Erlangen-Nürnberg

Corinna Reinhardt
Institute of Classical Archeology, Friedrich-Alexander-Universität Erlangen-Nürnberg

Vincent Christlein
Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg

Andreas Maier
Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg

Peter Bell
Institute for Art History, Friedrich-Alexander-Universität Erlangen-Nürnberg

Keywords: classical archaeology, computer vision, deep learning, convolutional neural networks, vase paintings
Topics: Europe, English, BCE-4th Century, 5th-14th Century, 15th-17th Century, artificial intelligence and machine learning, semantic analysis, Archaeology, Computer science

The understanding of iconography and visual narration in ancient imagery is one of the main foci in the field of Classical Archaeology, e.g. in Attic vase paintings of the fifth century B.C. In order to depict the situations and actions of a narrative as well as to characterise its protagonists, ancient Greek artists made use of a broad variety of often similar image elements [1]. The interaction and meaningful relationship of the protagonists is depicted with significant postures and gestures (schemata) in order to illustrate key aspects of the storyline [2, 3]. These schemes are not restricted to a certain iconography, so that visual links between different images occur. Being familiar with these relationships the ancient viewer could detect the specific narration and understand the meaning of the image. For example, the scheme of

leading the bride in Attic vase paintings is characterised by a significant leading-gesture (χεῖρ' ἐπὶ καρπῷ - hand on wrist / hand on hand) that relates bride and bridegroom in non-mythological wedding scenes, thereby expressing a hierarchy of the two figures in active or passive parts respectively. Both protagonists are connected in a communicative way defined by meaningful postures and gestures.Besides these descriptive schemes there are narrative elements, like the surrounding, specific attributes of gods or other objects, that contribute to the viewer's understanding of narrative contexts. These other characteristic attributes are needed in order to determine the depiction as a specific mythological one. Scenes of Helen and Menelaos (dressed as warrior and threatening her with weapons), thus, make use of the leading the bride scheme to show that Menelaos leads Helen home from Troy, like a bride. Only by combining the formal schemes and narrative elements is it possible to determine the depiction as a certain mythological story and to understand its cultural meaning.We, therefore, need to recognize the narrative elements of the pictorial stories and to decode their formal key structures by comparing a high number of ancient images and by discovering visual similarities. It is necessary to target the understanding of complex semantic relationships and cultural references in a mixed-initiative-interaction between researcher and machine, and involve computer vision as a fourth Bildwissenschaft [4].We approach this problem using state-of-the-art object recognition algorithms [5, 6] in computer vision. We present a novel approach using convolutional neural networks which maps the attributes occurring in Classical Archaeology into a deep representational space, through deep learning techniques. This representation maps the inherent variations in the representation of any particular attribute into a common space. A further novel method is developed that uses this representation space along with the context of the narrative for a better understanding of the depicted scene in the artwork. We also analyse the model performance with abstract classes where we combine multiple similar attributes into a single abstract class. This new approach has important applications, including the retrieval and contextual understanding of artworks [7] in Classical Archaeology.

References:

[1] Giuliani, L., 2003. Bild und Mythos. Geschichte der Bilderzählung in der griechischen Kunst, München: Beck 2003.

[2] McNiven, T. J., 1982. Gestures in Attic Vase Painting: Use and Meaning, 550-450 B.C., Diss. Ann Arbor: University of Michigan.

[3] Catoni, M. L., 2008. La comunicazione non verbale nella Grecia antica : gli schemata nella danza, nell'arte, nella vita, Turin: Bollati Boringhieri.

[4] Panofsky, E., 2006. Ikonographie und Ikonologie: Bildinterpretation nach dem Dreistufenmodell, Cologne: DuMont.

[5] Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[6] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[7] Garcia, N., & Vogiatzis, G. (2019). How to read paintings: Semantic art understanding with multi-modal retrieval. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11130 LNCS, 676-691 . https://doi.org/10.1007/978-3-030-11012-3_52

Reproducible Humanities Research Developing Extensible Databases for Recording "Messy" Categorisation, Annotation and Provenance Data

Melodee Beals
m.h.beals@lboro.ac.uk
Loughborough University, United Kingdom

Albert Meroño-Peñuela
Vrije Universiteit Amsterdam, Netherlands

Keywords: open data, databases, ontology, linked open data, multi-modal analysis
Topics: Global, English, Contemporary, database creation, management, and analysis, linked (open) data, Art history, History

Although the Digital Humanities is fundamentally interdisciplinary in nature, all humanities research questions require a degree of interdisciplinary thinking. History, for example, draws upon most other social sciences and humanities for obtaining and analysing source materials in different contexts. The multi-modal nature of these sources, the mixing of methodologies into bespoke, project-specific frameworks and the collaboration of researchers with overlapping but distinct interpretations all require a flexible workspace.

Moreover, growing calls for open research methods put pressure on humanities researchers to rethink how they document the provenance of their source materials as well as their interpretations.[1] Individual scholars often develop extensive, single-use taxonomies to categorise, encode and describe their conclusions; stored in a variety of document, spreadsheet and database systems, these are rarely disseminated and remain offline penumbra of the research process. Moreover, the prescriptive nature of out-of-the-box software may constrain the annotation process.[2] Larger collaborations may spend significant time developing extensive coding criteria resulting in over-fitted schema with little reusability or reach despite often herculean efforts of dissemination. Even when reusable, these schemas may require a degree of familiarity with the bespoke systems that makes them inaccessible to those outside the project. In order to overcome these difficulties, we have developed a highly extensible database development interface, Nisaba.[3]

Rather than prescribe a new database structure or encoding format, Nisaba was developed in order to accommodate a wide variety of source materials, encoding schema and dissemination formats. To achieve this, Nisaba leverages World Wide Consortium (W3C) standards[4] and Linked Data publishing practices,[5] which encourage the explicit provision and reuse of vocabulary terms. Written in Python 3.6 using TKinter, a cross-platform graphical user interface (Linux, OS, Windows), Nisaba functions as both an input and retrieval mechanism. Users input data including text transcriptions, images and [in the future] audio/visual files and apply user-created controlled-vocabularies, free-text annotations and an extensible selection of metadata. Once inputted, users create a segment (a selection of words, pixels or seconds of audio-visual information) and apply further metadata or annotations, allowing a single item to have multiple overlapping annotations using different schema by different users. In order to facilitate the documentation and exportation of data that is restricted or within copyright, the database encodes these segments by word number (text), or relative position (image), allowing precise locators without necessarily exporting the original materials. All data inputs are time-stamped and attached to individual user records, allowing for multiple researchers to annotate the same segments while maintaining unambiguous lines of provenance and allowing longitudinal use of the databases by multiple projects. Once inputted, the material can be retrieved through a simple browsing mechanism (controlled vocabulary) or by exporting layers of the data to non-proprietary formats, currently JSON or Turtle (RDF), allowing for deeply humanistic forms of knowledge representation in a format suitable for computational analysis.[6] This talk will demonstrate the use of Nisaba for various project types and provide guidance on how to develop an open, highly documented dataset to accompany humanities research.

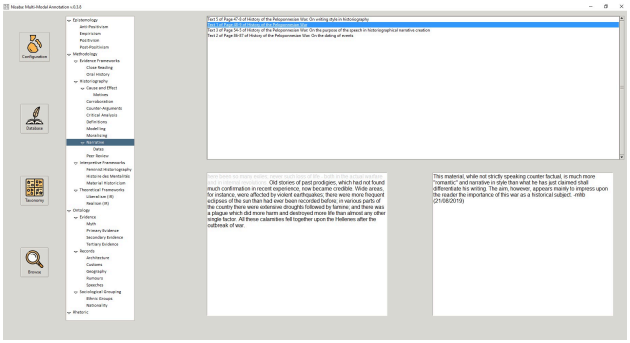


Figure 1:Text Segmentation

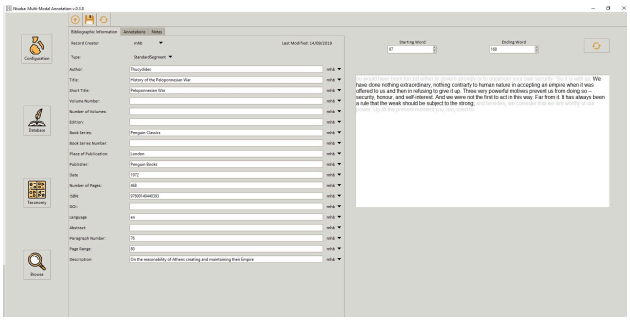


Figure 2: Browsing notes

[1] For more on the importance of evidence-interpretation-argumentation provenance, see D. M. Godden, "Arguing at Cross-Purposes: Discharging the Dialectical Obligations of the Coalescent Model of Argumentation", Argumentation, 17:2 (2013): 219-43.

[2] Noam Chomsky, Language and Mind (Cambridge: Cambridge University Press, 2006): 19.

[3] The open-source code is currently available at http://purl.org/nisaba

[4] See https://www.w3.org/TR/?tag=data

[5] Tom Heath and Christian Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1 (Morgan & Claypool: 2011).

[6] See Dominic Oldman, Martin Doerr and Stefan Gradmann, "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge" in Schreibmann, Siemens and Unsworth, eds. A New Companion to Digital Humanities (Oxford: Wiley-Blackwell, 2016): 251-73; David M. Berry and Anders Fagerjord, Digital Humanities: Knowledge and Critique in a Digital Age (Cambridge: Polity, 2017): 77.

Crossroads of Literature and History; Text Mining the Modern Persian Literature

Mehdy Sedaghat Payam
mspayam@umd.edu
University of Maryland, United States of America

Marjan Moosavi
University of Toronto, Canada

Keywords: Persian Literature, History, Modern, Text Mining
Topics: Asia, English, 20th Century, Contemporary, data modeling, text mining and analysis, Literary studies, South Asian studies

This paper analyzes the intersections of modern Persian literature and Iranian history through the text mining the modern Persian corpus.

Intersecting Digital Archives and GIS Technologies: Bringing IIIF Images to GIS Environments

Nung-yao Lin
nylin@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Shih-Pei Chen
Max Planck Institute for the History of Science, Germany

Keywords: GIS, IIIF, Georeferencing, Digital Archives
Topics: Asia, Global, English, 19th Century, 20th Century, Contemporary, software development, systems, analysis and methods, spatial & spatio-temporal analysis, modeling and visualization, Geography and geo-humanities, Library & information science

Since its publication in 2011, International Image Interoperability Framework (IIIF) has been widely adopted by archival institutions in particular. Among the image resources that are served on IIIF-compatible servers, many contain geographical references, whether images at or about certain places or historical maps for particular geographical regions. Due to the high cost of georeferencing, many archival institutions only digitize their resources and stop before georeferencing those images. Currently, IIIF does not address the issue of geo-referencing or geospatial metadata.

This paper describes a Web-based service that allows users to associate IIIF images with geographical coordinates in a much easier and more feasible way than georeferencing.

By associating IIIF images with geo-coordinates, it enables digitized cultural heritage images to be searched by their geographical locations rather than just by their metadata. It also enables them to be used with GIS technologies for geospatial observation and research.

Enriching Metadata for Irish Traditional Music at the American Folklife Center

Patrick Egan
patrick.egan@student.ucc.ie
Library of Congress, United States of America

Keywords: ethnomusicology, linked data, metadata
Topics: Global, Europe, English, North America, 20th Century, Contemporary, linked (open) data, metadata standards, systems, and methods, Musicology

In the past ten years, a number of digital projects have demonstrated the possibilities afforded by developing semantic ontologies, and a number of these ontologies have recently been published for music. Linked Jazz, for example, continues to demonstrate how discoveries can aid libraries and archives in re-imagining the representation of their holdings that contain music. The recently published

Linked Irish Traditional Music project (or LITMUS), is further testimony to value of Linked Data and the importance of building specific ontologies for Irish traditional music, song and dance.

In this presentation, I introduce Connections in Sound, a linked data case study that was focused on exploring Irish traditional music in audio collections from the American Folklife Center (AFC) at the Library of Congress. The project team worked closely with reference librarians to identify metadata from fieldworker and recordist documentation and transcribing audio data from a range of performance contexts. The resultant wide range of metadata allowed us to make connections between and unite metadata descriptions of audio material that had previously existed within separate sections of the archive. The case study demonstrates the particularities of a range of metadata, formats and fieldwork methods from material in collections that span the twentieth century at the AFC.

Connections in Sound pilots a number of digital methods such as data gathering and structuring to examine how material from disparate sources may be brought together and visualised with the digital re-combination of metadata. Connections in Sound progressed in an iterative, experimental way for a) gaining insight into the possibilities of utilizing current resources for digital semantic description within the genre and b) understanding limitations of these current resources as identified during data analysis and gathering.

This pilot seeks to contribute to discussions on what is specific about Irish traditional music that can inform how we might progress with Linked Data, for best practices, methods for analyzing, gathering and representing data. In so doing, Connections in Sound attempts to relay insight into understanding the opportunities and challenges that arise when developing linked data resources for Irish traditional music, and traditional music in general. It suggests potentials for continued research in this area and offers insight into advancements for Linked Data with traditional music in general.

The project is also supported up by a survey of musicians across North America who perform Irish traditional music in order to understand the nuances of their interaction with sound files from archives such as those within the American Folklife Center.

The Fight for National Language Rights in the USSR

Martha Brill Olcott
olcottm@msu.edu
Michigan State University, United States of America

Michael Downs
Michigan State University, United States of America

Ryan Lumbsden
Michigan State University, United States of America

Keywords: USSR, language rights, Kazakhstan, Kyrgyzstan, Moldova
Topics: Asia, Europe, English, 20th Century, crowdsourcing, digital archiving, History, Central/ Eastern European Studies

This presentation is part of "Collapse and Rebirth: A Living Archive on the End of the USSR and Afterward (1985-1994)." We have already developed the prototype website for the project, using a combination of two platforms (Esri Arcgis Storymaps, and Omeka S.The site will support public commentary by eyewitnesses, suggestions for additional events to be featured, and offers of archival material for curation and inclusion. It will include bibliographic entries for each event, and will provide links to other archival materials and digital resources on the topic that are available on the internet. We are proposing a lightening talk by MSU faculty and student team members at DH2020 introducing second phase "stories" on the political movements and protests that were key to getting rights of national language usage and status restored during the period 1987-1989 in several republics and autonomous regions of the USSR.

Performer communities within a hyperconnected company network

Antonio Jimenez-Mavillard
antonio.jimenez-mavillard@cssd.ac.uk

Royal Central School of Speech and Drama - University of London, UK

Kate Elswit
Royal Central School of Speech and Drama - University of London, UK

Harmony Bench
Ohio State University, USA

Keywords: dance, hyperconnectivity, network analysis, sankey diagram, movement community
Topics: South America, Europe, English, North America, 20th Century, network analysis and graphs theory and application, spatial & spatio-temporal analysis, modeling and visualization, African and African American Studies, Performance Studies: Dance, Theatre

This lightning talk presents our approach to identifying and representing what constitutes movement communities in the project Dunham's Data: Katherine Dunham and Digital Methods for Dance Historical Inquiry. The project uses the case study of Dunham's company, in which almost 200 dancers, drummers, and singers performed across the world for thirty years. The challenge this case poses is how to discover sub-communities among a hyperconnected community of artists who all worked together over time. Based on the concept of the check-in, which explicitly states the presence of performers on a specific date and location, we created a flow diagram and a community network that visualize the performers' individual paths through this hyperconnected network as well as the temporal sub-communities that these form. We analyzed the hyperconnectivity phenomenon from these two perspectives, in an effort to understand the complexity of this highly interconnected network.

Creating a Meaningful Genre Schema and Metadata using IMDb data for a Large-Scale Digital Humanities Project in Media Studies

Cindy Conaway
cindy.conaway@esc.edu
SUNY-Empire State College, United States of America

Diane Shichtman
diane.shichtman@esc.edu
SUNY-Empire State College, United States of America

Keywords: media studies, television studies, database, social network theory
Topics: English, North America, 20th Century, Contemporary, annotation structures, systems, and methods, metadata standards, systems, and methods, Film and cinema arts studies, Media studies

We are currently engaged in a long-term DH project examining the social networks of TV and film actors and crews across more than 32,500 media items they worked on, starting in 1938. The primary source is the Internet Movie Database (IMDb). IMDb is one of the most robust databases available and provides free downloadable data about the actors and crew that worked on various media items, including TV, movies, and video games. However, it is problematic in a number of ways, as we presented at DH2018. IMDb appears to be under-researched as a source for media studies in DH, with many scholars focusing on fan activity, or on film actors and directors only, with most not questioning IMDb genres. This presentation will explore the contrast between IMDb's methods and the schema we have had to create. It also discusses some of the challenges we faced, and how we think this could aid other researchers creating public digital humanities databases.

Steve Neale writes in Genre and Contemporary Hollywood, "genres can be approached from the point of view of the industry and its infrastructure, from the point of view of their aesthetic traditions, from the point of view of the broader socio-cultural environment upon which they draw and into which they feed, and from the point of view of audience understanding and response." We must consider all of these concepts as our work encompasses not only Hollywood film, the subject of most genre analysis, but media of every type including much that is obscure or forgotten. IMDb's genre methodology was inadequate, so we turned to industry-focused websites to enhance our understanding.

Other schemas use very vague macro descriptors, or idiosyncratic descriptors allowing a media item to be included in multiple "lists." A movie in the Library of Congress is simply Comedy, Drama, Action, etc. AFI divides films into "best" but also "most

thrilling" (including action, horror, and adventure). The Telegraph writes that Netflix's "genres, based on a complicated algorithm that uses reams of data about users' viewing habits . . . number in the tens of thousands" including commonly-accepted genres like "Action" but also "Family Watch Together TV." We attempted to taxonomize, expanding and limiting the options

available to encompass the full range of programming, but not get so esoteric that they cannot be clustered together and measured. We have created a vocabulary/schema to allow for clear communication as part of Public Humanities. We cannot, for example, talk about the representation of women in a particular subgenre if we do not have a shared understanding of it. If other scholars also use this schema (or work with us to adapt it), each media item can be described in a way that allows for effective and relatively consistent coding by multiple scholars.

Genre on IMDb is handled in ways that are not useful for analysis because terms are used in inconsistent ways. There's not enough inter-rater reliability, the tags are misleading, and scholars do not all agree on how to use them. As Deb Verhoeven states in "Mapping the Movies," a project like her team's "only works if the existing data collection is both sufficiently comprehensive and thoroughly reliable, since it will have to be accepted by all partners" (Verhoeven). There is little consistent agreement on genre among scholars or fans, as borne out in genre theory. What IMDb designates as "genres" actually combines traditional genres, subgenres, and target audience categories. As IMDb allows those who enter the data to select any number of these terms, and many fans enjoy labeling media items to fit multiple lists, it becomes impossible to analyze using IMDb's categories. In large part this is because the database relies heavily on users: sometimes cast/crew members, agents, producers, and fans, for its data and for much editing. As Wasserman et al., point out, "Although user editing allows a reference website such as IMDb to be up-to-date, it diffuses the responsibility for fact-checking, leading to greater uncertainty about accuracy and objectivity of information" (Wasserman).

It has taken significant additional research and reorganization to use the data effectively because, as media researchers Marsden, et al. explain, there is not enough agreement about metadata. While most people can tell a Western from Science Fiction, IMDb makes it more difficult to deal with hybrid genres such as Dramedies or Family movies, or where a particular movie or show combines genres, such superimposing Western generic concepts into Science Fiction, or an Action/Adventure movie with a strong romantic plot. The schemas used by the Library of Congress, Netflix, Amazon and others are too reductive or imprecise for our purposes.

Therefore, we not only had to create a taxonomy with a variety of categories, including subjects, styles, modes, and purposes, but our own concise definitions for these categories. We will share parts of our schema in this presentation.

Conaway, Cindy and Diane Shichtman. "Seinfeld at The Nexus of the Universe: Using IMDb Data and Social Network Theory to Create a Digital Humanities Project." DH2018. Mexico City. MX. June 24-30, 2018.

Lutter, Mark. "Do women suffer from network closure? The moderating effect of social capital on gender inequality in a project-based labor market, 1929 to 2010." American Sociological Review 80.2 (2015): 329-358.

Lutter, Mark. "Creative success and network embeddedness: Explaining critical recognition of film directors in Hollywood, 1900-2010." Max Planck Institute for the Study of Societies. (2014).

Marsden, Alan et al. "Tools for Searching, Annotation and Analysis of Speech, Music, Film and Video--A Survey." Literary and Linguistic Computing 22.4 (2007): 469-488. Web.

Neale, Stephen. Genre and Contemporary Hollywood. British Film Inst, 2002.

Rossman, Gabriel, Nicole Esparza, and Phillip Bonacich. "I'd Like to Thank The Academy, Team Spillovers, and Network Centrality." American Sociological Review 75.1 (2010): 31-51.

Titcomb, James. "Netflix Codes: The Secret Numbers that Unlock Thousands of Hidden Films and TV Shows." The Telegraph, -12-20 2017, Web. Jan 3, 2020 .

Verhoeven, Deb, Kate Bowles, and Colin Arrowsmith. "Mapping the Movies." Digital Tools in Media Studies. 1. 2015.

Wasserman, Max, et al. "Correlations between User Voting Data, Budget, and Box Office for Films in the Internet Movie Database." Journal of the Association for Information Science & Technology 66.4 (2015): 858-68.

DIGITAL HUMANITIES SCHOLARSHIP IN AFRICA: Prospects and Challenges

Augustine Farinola
aaf936@student.bham.ac.uk
University of Birmingham, United Kingdom

Keywords: Digital Humanities, African, Researchers, Tools
Topics: Africa, English, 20th Century, Contemporary, software development, systems, analysis and methods, user experience design and analysis, African and African American Studies, Humanities computing

This research addresses the issues surrounding the low level of Digital Humanities (DH) technological consciousness among students and academics in the humanities discipline in Africa (Nigeria). The study, using online questionnaires, shows that despite the wide acceptance of DH Technological tools among some African scholars in the humanities, there are still challenges experienced by these Scholars in the course of using some of these DH tools to capture African realities. These difficulties include low level of training for users of DH technologies in Africa, as well as the designers' failure to optimize those tools for use in the analysis of data, texts, and images extracted in Africa. To address these constraints, this research enjoins African scholars to come up with epistemological and ontological frameworks that would aid software developers in creating tools which capture the unique aspects of African history, techne, culture, philosophy and tradition.

Comparative analysis of rhythms in Japanese folk songs

Akihiro Kawase
kawase@dh.doshisha.ac.jp
Doshisha University, Japan

Keywords: rhythm, folk music, nPVI, GIS
Topics: Asia, English, 18th Century, 19th Century, 20th Century, cultural analytics, music and sound digitization, encoding, and analysis, Cultural studies, Musicology

In this study, we applied nPVI to rhythm analysis. nPVI is an index proposed to quantify the difference between stress and syllable beats in linguistics and has recently been used for the comparative analysis of rhythms in Western music history. We digitized a total of 5,484 songs published in all nine volumes of "Nihon Min'yo Taikan" (Anthology of Japanese Folk Songs). The difference in the agogic contrasts was clarified with the folk songs from each region. In addition, we calculated the latitude and longitude of the place from where each song was recorded and created a database in a format that can be used for GIS. By converting the data into mesh sections and integrating the nPVI values, the difference in the rhythm leaps of each song in each region was visually clarified.

Encoding Belief: why digital has become the new location of salvation and destruction.

David Stewart White
david.white@arts.ac.uk
University of the Arts London, United Kingdom

Keywords: digital, belief, religion, sacred
Topics: Global, English, 15th-17th Century, Contemporary, cultural analytics, rhetorical analysis, History of science, Theology and religious studies

Extending the theme of the religion of technology and data gathered from a textual analysis of technology reporting this presentation seeks feedback and discussion on the proposal that digital is the location of a new form of implicit religion and commonly engaged with as sacred.

Revenge of the Silk. Elaborating and using a CIDOC CRM extension a case study on silk-related artefacts

Marie Puren
marie.puren@cnrs.fr
Laboratoire de recherche historique Rhone-Alpes / CNRS, France

Pierre Vernus
pierre.vernus@ish-lyon.cnrs.fr

Laboratoire de recherche historique Rhone-Alpes / Université Lumière Lyon 2, France

Keywords: CIDOC CRM, history of technics, textile, linked data
Topics: Europe, English, 15th-17th Century, 18th Century, 19th Century, data modeling, data, object, and artefact preservation, History

If you are interested by silk, you would like to access the description of several pieces of textiles, to compare and study them. But it can be difficult to study those that are not preserved in the same museum : they can be described in languages you don't know, they can be hard to find, and you cannot obtain these results via a single query, and in a single environment.

To make it easier to find information on silk-related artefacts, the future SILKNOW online platform proposes to accommodate these digital data from different sources by using a common language, CIDOC Conceptual Reference Model or CIDOC CRM.

CIDOC CRM is a core ontology with more specialist extensions. Using a few examples coming from Cultural Heritage Institutions' catalogs and databases, we present the CRM extension we are currently elaborating for this purpose. We also ask for feedback on a general workflow we are currently designing to help those who are interested in using CIDOC CRM and/or are in need to develop CRM extensions. This workflow using the Ontology Management Environment (OntoME) could be re-used by research projects dealing with the same kind of issues, and wishing to find a solution to widely disseminate digital Cultural Heritage data.

References

CIDOC CRM Homepage, http://www.cidoc-crm.org/. Last accessed 07 October 2019.

Le Boeuf, P., Doerr, M., Ore, C., Stead, S.: Definition of the CIDOC Conceptual Reference Model, Produced by the ICOM/CIDOC Documentation Standards Group, Continued by the CIDOC CRM Special Interest Group Version 6.2.1, October 2015, http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.1.pdf. Last accessed 07 October 2019.

Ontology Management Environment (OntoME), http://ontome.dataforhistory.org/. Last accessed 14 October 2019

SILKNOW Homepage, http://silknow.org/. Last accessed 07 October 2019.

CLARIN for DH Scholars

Leon Wessels
l.c.wessels@uu.nl
Utrecht University / CLARIN ERIC, Netherlands, The

Keywords: CLARIN, research infrastructure, fair principles, language resources
Topics: Europe, English, Contemporary, digital research infrastructures development and analysis, Humanities computing, Linguistics

The increased application of data-driven approaches has been a game changer in the Humanities. By using large quantities of research data and various tools to process and analyse these data, Digital Humanities scholars can address questions that were previously considered too complicated or time-consuming to answer. But the development of the DH field has also reshaped the needs of researchers. DH scholars desire increasingly larger, sufficiently annotated sets of research data and advanced tools to process them. This abstract introduces a number of services offered by CLARIN, the Common Language Resources and Technology Infrastructure, that are particularly interesting for the DH community.

As a European Research Infrastructure Consortium (ERIC) established by the European Strategy Forum on Research Infrastructures (ESFRI), CLARIN is a non-commercial Research Infrastructure providing single sign-on access to natural language resources and tools free of charge for all academic researchers.[1] Countries or regions can join CLARIN as member, observer, or third party. As of June 2020, the CLARIN consortium consists of 21 full members, three observers, and a third party, covering a total of 25 participating countries (figure 1). Within each national consortium a number of institutes (universities, academies, research institutes, museums, archives, etc.) contribute resources, tools, and knowledge to the CLARIN infrastructure.

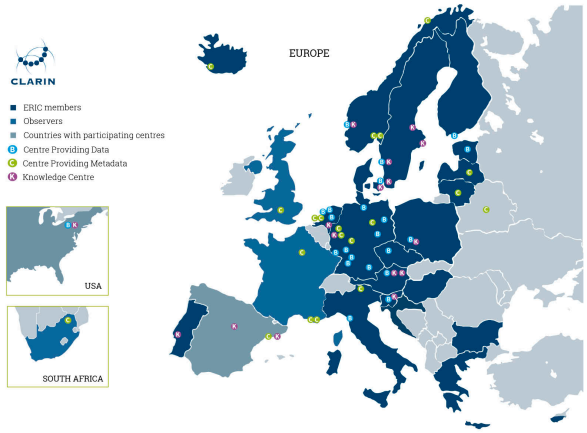


Figure 1: This map shows countries and regions participating in CLARIN as member, observer, or third party, and various types of CLARIN centres.

CLARIN aims to cater the needs of the entire DH community. Clearly, these needs differ widely. Some DH researchers, for example, like to collect, curate, and deposit their own data sets and to develop their own tools, while others prefer to focus on solving a piece of the puzzle of the various aspects of human society and culture, leaving the technical development to others. Surely, there is no such thing a one-size-fits-all solution to cater to the needs of all different types of DH researchers. However, a number of developments would benefit the DH community at large. For instance, making resources and tools as sustainable, openly available, interoperable, and easily findable as possible. From the outset, it has been CLARIN's strategy to make its resources and tools available according to the later defined FAIR data principles: Findable, Accessible, Interoperable, and Reusable.[2]CLARIN is a distributed infrastructure. Researchers can deposit their resources in one of the certified CLARIN centres offering open repositories, each devoted to a specific research field. The metadata of each deposit gets automatically harvested by the Virtual Language Observatory (VLO, https://vlo.clarin.eu/), a faceted search engine allowing everyone to find resources deposited in all of the CLARIN repositories. [3] Currently, the VLO gives access to over 900,000 records. Once a relevant resource has been found it can be immediately processed by a number of analytical tools for part-of-speech tagging, distant reading, topic modelling, named entity recognition, machine translation, and various other tasks using the CLARIN Language Resource Switchboard (http://switchboard.clarin.eu/). The Switchboard takes into account features like modality, format, and language to match resources to tools.[4] In addition to a full inventory of all resources made available through the various CLARIN repositories, CLARIN also provides user-friendly overviews of key resources. These overviews of CLARIN Resource Families (https://www.clarin.eu/resource-families) provide information on the availability, language, size, annotation, and license of the resources. Currently, ten corpora families, five families of lexical resources, and three tool families are offered. The corpora families include parliamentary corpora, newspaper corpora, literary corpora, spoken corpora, and computer-mediated communication corpora (e.g. social media posts). The overviews are initiated based on the input of domain experts across the world and continue to be manually curated.[5]But CLARIN offers more than just data and tools. Knowledge is a key component of the CLARIN infrastructure. A coordinated system of Knowledge centres provides knowledge and expertise to researchers. Each Knowledge centre has its own specific area of expertise, such as data management, language learning analysis, speech analysis, treebanking, and several languages. [6] Several funding instruments allow researchers, teachers, and developers to collaborate and to teach each other, stimulating cross-country and cross-disciplinary collaboration.[7] Dozens of recorded presentations, tutorials, discussions, and other videos have been made available on VideoLectures.NET (http://videolectures.net/clarin/).

Researchers, teachers, developers, citizen-scientists, policy makers, politicians, journalists, and other people interested in getting to know more about CLARIN are invited to have a look at the CLARIN Value Proposition (https://www.clarin.eu/value-proposition). Those who would like to know what is going on in their national CLARIN consortium can have a look at the list of participating consortia (https://www.clarin.eu/content/participating-consortia) or at the Tour de CLARIN webpage (https://www.clarin.eu/Tour-de-CLARIN), an ongoing initiative to highlight the activities in national consortia and Knowledge centres.

References

1. Erhard Hinrichs and Steven Krauwer, The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 1525-1531. Handle: http://dspace.library.uu.nl/handle/1874/307981; Sheena Bassett, Leon Wessels, Steven Krauwer, Bente Maegaard, Hella Hollander, Femmy Admiraal, Laurent Romary, and Frank Uiterwaal, Connecting the Humanities through

Research Infrastructures. Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019). URL: https://hal.inria.fr/hal-02047512/document; Franciska de Jong, CLARIN - Infrastructural Support for Impact through the Study of Language as Social and Cultural Data. In: Bente Maegaard and Riccardo Pozzo eds., Stay Tuned to the Future. Impact of the Research Infrastructures for Social Sciences and Humanities, Leo S. Olschki Editore, 2019, pp. 121-129. URL: https://www.clarin.eu/sites/default/files/STAY_TUNED_TO_THE_FUTURE.pdf.

2. Franciska de Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer, and Dieter Van Uytvanck, CLARIN: Towards FAIR and Reponsible Data Science Using Language Resources. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 3259-3264. URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/575.pdf. Up-to-date information concerning CLARIN and FAIR Data can be found on: https://www.clarin.eu/fair.

3. Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen, Semantic metadata mapping in practice: The Virtual Language Observatory. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), pp. 1029-1034. Handle: http://hdl.handle.net/11858/00-001M-0000-0014-744B-B; Twan Goosen and Thomas Eckart, Virtual Language Observatory 3.0: What's New? Paper presented at the CLARIN Annual Conference 2014. URL: http://www.clarin.eu/sites/default/files/cac2014_submission_2_0.pdf. Up-to-date information concerning the Virtual Language Observatory can be found on: https://vlo.clarin.eu/help.

4. Claus Zinn, The CLARIN Language Resource Switchboard. Proceedings of the CLARIN Annual Conference 2016. URL: https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf. Up-to-date information concerning the CLARIN Language Resource Switchboard can be found on: https://switchboard.clarin.eu/help.

5. Darja Fišer, Jakob Lenardi?, Tomaž Erjavec, CLARIN's Key Resource Families. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 1320-1325. URL: http://www.lrec-conf.org/proceedings/lrec2018/pdf/829.pdf. Up-to-date information concerning the CLARIN Resource Families can be found on: https://www.clarin.eu/resource-families.

6. An overview of Knowledge centres can be found on: https://www.clarin.eu/content/knowledge-centres.

7. An overview of CLARIN's funding instruments can be found on: https://www.clarin.eu/funding.

Computational Analysis of Emotions and Topics in Survivor Interviews about Nanking Massacre

Lu Xiao
lxiao04@syr.edu
Syracuse University, United States of America

Jianyi Liu
Syracuse University, United States of America

Wenchao Zhai
Syracuse University, United States of America

Liangqin Jiang
Nanjing University, Nanjing, China

Keywords: emotion analysis, topic analysis, Nanking Massacre
Topics: Asia, English, 20th Century, natural language processing, text mining and analysis, History, Humanities computing

In survivor interviews, often the interviewee's deep feelings or attitudes towards their painful memories or experiences is the focus of the conversation. Past research has shown that sensitive interviews such as survivor interviews affect the emotional state of the interviewee. In this project, we explored this aspect in the survivor interviews about Nanking Massacre. Specifically, we examined the major types of interviewees' emotions (e.g., fear, sadness, fear). in the interviews, where they occur during the interview turns, and the interviewers' emotions near those turns. In addition, we applied topic modelling analysis to explore the number of topics in these interviews and how different topics are related to the different emotions.

Does Culture Still Matter? A Cross-Cultural Perspective on Human-Robot Interaction and Artificial Intelligence

Serdar Tuncer
stuncer2@uottawa.ca
University of Ottawa, Canada

Boulou Ebanda De B'Beri
University of Ottawa, Canada

Keywords: Artificial Intelligence (AI), Human Robot Interaction (HRI), Negative Attitudes Towards Robots Scale (NARS), 6D Cultural Model, Cross-Cultural Perspective
Topics: Comparative (2 or more geographical areas), English, Contemporary, artificial intelligence and machine learning, public humanities collaborations and methods, Cultural studies, Humanities computing

Humanity is facing a turning point in its history. Artificial Intelligence (AI) plays a crucial role not only in computer science and robotics but also in the communication between machines and machines, and between humans and machines. With the rise of this new interactive space, it is becoming ever more critical to ask what role intercultural communications will play in this complex correlation. Does culture still matter in the era of AI?

To explore this question, we mobilize Geert Hofstede's 6D cultural model (2010, 2020), one of the most comprehensive models for describing a country's characteristics and apply it to the Negative Attitudes Towards Robots Scale (NARS) (Nomura, Kanda, & Suzuki, 2006), a study which has been conducted in six countries (China, Germany, Japan, Mexico, Netherlands, United States) to measure human attitude towards robots within the cross-cultural perspective (Bartneck et al, 2005). In this paper, we understand culture as Hofstede's 'National Culture' which he defines as "the collective programming of the mind distinguishing the members of one group or category of people from others" (Hofstede Insights 2020). The NARS scale provides a foundation that has been used in several studies to measure human attitudes towards humanoid robots (Riek et al, 2010). Together these two approaches provide a cross-cultural framework for analyzing an emerging era of communications in the context of Human-Robot Interaction (HRI).

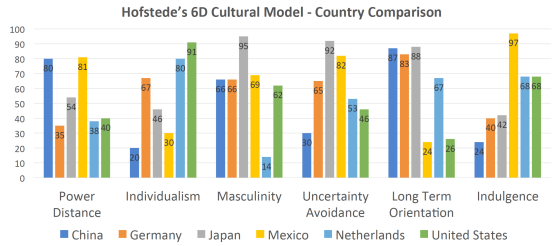


Figure 1: Hofstede's 6D Cultural Model - Country Comparison

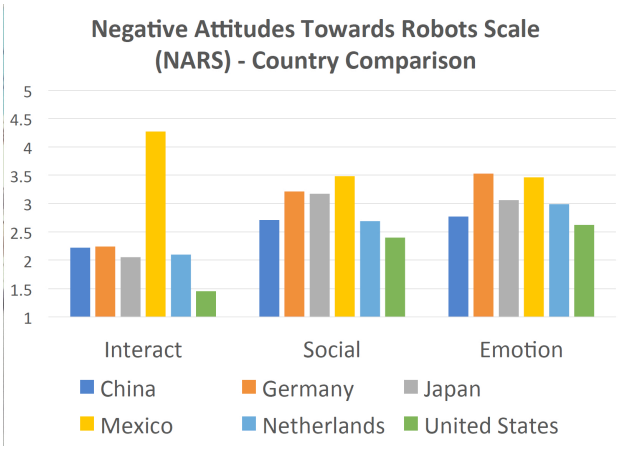


Figure 2: Negative Attitudes Towards Robots Scale (NARS) - Country Comparison

Our initial findings show there is a correlation between a country's 'Individualism', 'Power Distance' and its attitude towards HRI. For instance, highly individual societies with less power inequality, like the United States, have more positive attitudes towards HRI when compared to collectivist cultures, like Mexico. One exception is China; this could be due to the country's high level of technology manufacturing involving industrial robots. Another of our findings is that countries that are 'Long-term Oriented', such as Netherlands and Germany, are less negative in regard to HRI, compared to those who focus on the short term, like Mexico. The exception here is the United States; this could be that despite the States' 'Short-term orientation', it has a long history of interacting with

technology and robots. A tendency also exists between a country's 'Uncertainty Avoidance' and their attitude towards the social influence of robots, with a greater desire to avoid uncertainty correlating to more negative attitudes toward robots' social influence. In terms of attitude towards 'Emotions' in interaction with robots, a correlation with Hofstede's 6D cultural model was not identified.

As a pilot study combining Hofstede's 6D cultural model and NARS, our aim with this study is to identify possible empirical correlations that can then be supported through a larger multi-country study evaluating AI and HRI through a cross-cultural lens. Our critical comparative analysis illustrates that, though technological development cannot be contained within a country's territorial boundaries, culture is a crucial variable in HRI. A cultural and qualitative perspective is needed in future studies of HRI to understand what role intercultural communications will play in this complex interaction.

REFERENCES:

Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., & Kennsuke, K. (2005). A cross-cultural study on attitudes towards robots. Proceedings of the HCI International, Las Vegas.

Country Comparison - Hofstede Insights. (2020). Hofstede Insights. Retrieved 1 May 2018, from https://www.hofstedeinsights.com/country-comparison.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and organizations. New York, NY: McGraw-Hill.

Nomura T, Kanda T, Suzuki T. (2006). Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. AI Soc 20(2):138-150.

Riek, L. D., Mavridis, N., Antali, S., Darmaki, N., Ahmed, Z., Al-Neyadi, M., et al. (2010). "Ibn sina steps out: Exploring Arabic attitudes toward humanoid robots," in Proceedings of the 2nd International Symposium on New Frontiers in Human-robot Interaction, AISB, Leicester (Leicester), 1.

INVESTIGATING THE USE OF SOCIAL MEDIA FOR HEALTH CARE ADVOCACY IN NIGERIA: A DIGITAL MEDICAL HUMANITIES APPROACH

Tunde Ope-Davies
bopeibi@unilag.edu.ng
CENTRE FOR DIGITAL HUMANITIES, UNIVERSITY OF LAGOS, Nigeria

Mojisola Shopide
CENTRE FOR DIGITAL HUMANITIES, UNIVERSITY OF LAGOS, Nigeria

Keywords: Technology, Health Humanities, New Media
Topics: Africa, English, Contemporary, digital activism and advocacy, social media analysis and methods, Communication studies, Humanities computing

This study applies computer-mediated communication models (Thorne, 2008; Herring, 2004) to investigate the role of digital technologies for health communication and person-centred care. It relies on Kirsten Ostherr's (2019) insight on digital health humanities "as an emerging research method that blends critical analysis with the use of computational tools to explore research questions related to digital information and communication technologies (ICTs) in health care."

The study focuses on the use of WhatsApp for the broadcast of health-care information and wellness advocacy among citizens. It highlights how new media technologies are redefining health information and medical discursive engagement in Nigeria. It is suggested that technology will help to drive narrative medicine which will include the development of apps and patients online narratives to enable future patients benefit from shared data (Ostherr, ibid.)

Abusing the Concept of Normalization for Better Collation Results (and Profit)

Tara Lee Andrews
tara.andrews@univie.ac.at
University of Vienna, Austria

Keywords: text collation, normalization, cheap tricks
Topics: Global, English, 5th-14th Century, Contemporary, data modeling, scholarly editing

and editions development, analysis, and methods, Humanities computing, Philology

In this lightning talk I will present a counter-intuitive use of reading normalization to improve the results of automated collation.

Prosopography in the digital era? Experiments in managing data from a Historical and Biographical dictionary

Jean S. e Spritzer
jean.spritzer@fgv.br
Fundação Getulio Vargas, Brazil

Keywords: Prosopogaphy, Biography, Text Minin, Brazilian Politics
Topics: South America, English, 20th Century, Contemporary, digital biography, personography, and prosopography, text mining and analysis, Political science, Sociology

When we refer to a large data amount, even if the sources are reliable, techniques of data mining and structuring still presenting as a challenge. Within this agenda, this presentation explores the uses of digital methodologies - specifically, text mining - in a historical dictionary composed of almost 8 thousand entries related to Brazilian Politics and from which we propose to compile a prosopographic database.

The visual digital humanities – topics, researchers and cultures

Sander Muenster
sander.muenster@uni-jena.de
FSU Jena, Germany

Keywords: bibliometrics, survey, scholarly communities
Topics: Global, Europe, English, Contemporary, bibliographic analysis, Library & information science

The data foci of digital humanities are texts, images and objects. While the use of digital methods in the text-oriented disciplines is currently widely established and standardized, a scope of digital methods related to images and other visual objects and basing on vision rather than close reading remains - despite various attempts - essentially undiscovered. Against this background, three areas of usage of visual oriented methods and approaches are of interest for our investigation: (a) Scholars working in visual digital humanities, (b) Fields of research, topics and methods used by these scholars, (c) Institutionalization & disciplinary culture of these scholars. Investigations were done via 15 expert interviews with researchers in London and 6 interviews in Los Angeles as well as via two surveys with more than 900 participants each. Key findings are about disciplinary backgrounds and about how scholars enter the digital humanities as well as topics and international collaborations and project funding.

Transatlantic Cultures: a digital humanities project for a connected history of the Atlantic world (18th-21st centuries)

Anaïs Fléchet
CHCSC Laboratory, UVSQ, University of Paris-Saclay

Olivier Compagnon
IHEAL-CREDA, Sorbonne Nouvelle Paris 3 University

Gabriela Pellegrino Soares
FFLCH, University of Sao Paulo

Cyrille Suire
cyrille.suire@uvsq.fr
CHCSC Laboratory, UVSQ, University of Paris-Saclay

Elina Djebbari
IHEAL-CREDA, Sorbonne Nouvelle Paris 3 University

Keywords: connected history, atlantic space, cultural history, circulations
Topics: Comparative (2 or more geographical areas), English, 18th Century, 19th Century, 20th Century, cultural analytics, digital publishing projects, systems, and methods, Cultural studies, History

This lightning talk will present Transatlantic Cultures, a joint research project of faculty from universities in France, Brazil, the United States, Mexico, Senegal, and other countries. Its purpose is to produce a Digital Platform for Transatlantic Cultural History, edited in four languages which aims to analyze the cultural dynamics of the Atlantic region and its central role in the contemporary process of globalization. The online digital platform will present to the public a connected history of the Atlantic space since the end of the 18th century through a series of digitally enhanced analytical essays exploring cultural relations and exchanges between Europe, Africa, and the Americas from multidisciplinary perspectives.

Digital Approaches to Name Disambiguation of Chinese Historical Figures

Lik Hang Tsui
lhtsui@cityu.edu.hk
City University of Hong Kong, Hong Kong S.A.R. (China)

Keywords: Name disambiguation, Chinese history, Biographical data, Gazetteers, Personal names
Topics: Asia, Global, English, 5th-14th Century, 15th-17th Century, 18th Century, digital biography, personography, and prosopography, Asian studies, History

When integrating biographical data extracted from 2,000+ local gazetteers into the China Biographical Database (CBDB), we need to identify and link records of the same person--the act of "disambiguating" them. Traditional Chinese naming customs pose big challenges to this, especially for the gazetteer dataset containing 0.12 million records and 90k unique names of imperial government officials. Also, useful variables are missing in numerous entries in these gazetteers. My presentation analyzes solutions to disambiguating identical personal names in Chinese script. First, we identified the individuals who repeatedly took official posts in the same locality. Then, we cross-tabulated the overlap of content in multiple gazetteers. Finally, we corroborated the remaining data with external datasets e.g. CGED-Q of the Lee-Campbell research group. Through doing so we have disambiguated 51k personal names with optimal precision. Such task is only possible if done digitally. The techniques explored in this study will also be useful for disambiguation and Named Entity Recognition of other large-scale unstructured data in non-Latin script.

Diffusion and Change of the Romantic Concept in Historical Newspapers

Tinghui Duan
tinghui.duan@uni-jena.de
Graduate School "The Romantic Model" at University of Jena Jena University Language & Information Engineering Lab (JULIE Lab)

Udo Hahn
Jena University Language & Information Engineering Lab (JULIE Lab)

Keywords: Romanticism, Historical Newspaper, Cultural Diffusion, Named Entity Recognition, Natural Language Processing
Topics: Comparative (2 or more geographical areas), Europe, English, North America, 18th Century, 19th Century, 20th Century, natural language processing, spatial & spatio-temporal analysis, modeling and visualization, Literary studies, Philology

Our in-progress project analyzes digitized historical newspapers at a very large scale. It aims to: 1) trace the diffusion of the Romantic concept from its origin in German Romanticism to other countries, other languages, and other time periods; 2) unveil how the Romantic concept has changed in the past two hundred years.

Starting from the "Allgemeine Literatur-Zeitung" (General Literature Gazette, 1785-1849) - an authoritative textual resource for research on German Romanticism, we have extracted named entities that represent the Romantic concept by using state-of-the-art NER methods. These NEs will then be searched in larger newspaper corpora.

Since the Romantic concept has moved beyond its original context and survived without mention of its representative figures (NEs), we also analyze the usage change of its related words and word groups by using sophisticated natural language processing (NLP) methods.

Rhetorical interpretation of stylometry methods in disciplinary writing: a question of hermeneutics

Hannah Ringler
hringler@andrew.cmu.edu
Carnegie Mellon University, United States of America

Keywords: stylometry, rhetorical interpretation, hermeneutics, function words, disciplines
Topics: Global, English, Contemporary, attribution studies and stylometric analysis, rhetorical analysis, Linguistics, Literacy, composition, and creative writing

This work focuses on investigating why function word frequency, a commonly used metric in stylometry research, performs so well in distinguishing academic disciplinary writing. How does function word frequency fit with and enrich our broader theoretical conceptions of disciplines as social constructs with particular "ways of knowing and doing" (Carter, 2007) and as performing kinds of "social action" (Miller, 1984)? More generally speaking, how can focusing on frequent language-level features that corpus work reveals help us to better articulate how rhetorical action is accomplished through specific linguistic resources? I focus my analysis on function words that contribute to the difference between the humanities and sciences broadly, and ask how those words support particular types of textual moves that undergird the production of knowledge in the disciplines.

Cyborg Public Humanities: The Digital-Material Exhibit and the Classroom

Deanna Stover
d.m.stover@tamu.edu
Texas A&M University, United States of America

Michaela Baca
michaela510@tamu.edu
Texas A&M University, United States of America

Keywords: public humanities; student-centered learning; digital exhibits; material history
Topics: Europe, English, 3D printing, critical making, 15th-17th Century, 19th Century, Contemporary, public humanities collaborations and methods, Book and print history, Literary studies

This short presentation discusses how to incorporate a student-led, interdisciplinary, public-facing approach in the classroom. We use an example from our own experience: a "cyborg" public humanities project that blended the digital with the material, the academic with the public, and the visual with the written. Students in two literature classrooms used the science fiction collection at Texas A&M University to create digital posters accessible through QR codes and physical posters about and related to Frankenstein to celebrate the 200th anniversary of that iconic novel. The combination of digital and material components led to an interactive exhibit in Cushing Memorial Library--a polyvocal display that granted ownership to the expertise of our students. Hands-on public projects serve both students and teachers, and when we think of "cyborg" projects--neither purely digital nor material--we expand the value of the humanities outside of the traditional classroom.

Women Personal Archives: the digitization of CPDOC's feminine collection and the relevance of women political action in contemporary Brazil

Carolina Alves
carolina.alves@fgv.br
Getulio Vargas Foundation (FGV), Brazil

Keywords: Personal Archives, Women, Memory, History, Digitization
Topics: South America, English, 20th Century, Contemporary, digital archiving, digitization (2D & 3D), Gender and sexuality studies, Sociology

The increasing political participation of women contrasts with historical data showing a shy feminine participation in Brazil's public affairs. This shyness, which is closely related to the reduced presence of women in hegemonic historical narratives, leaves distinguishing marks in Brazilian archives. The lack of women-collected documents in archival institutions should be considered under a light in which the presence of women in the public life is silenced. The digitization of CPDOC's feminine collection have the purpose of increasing the amount of historical sources made available online. Of the over 2.7 million pages of archives in the institution, more than 900 thousand have been digitized and made available online. These documents can be considered the largest available and accessible primary source of Brazilian history on the internet. The women archives proposed for digitization will increase the academic relevance of this digital collection, especially considering its specificity.

Pricing Opinions and Tastes: The Market for « Primitive » Artefacts through Sentiment Analysis

Léa Saint-Raymond
lea.saint.raymond@gmail.com
ENS & CNRS, France

Thierry Poibeau
thierry.poibeau@ens.fr
ENS & CNRS, France

Keywords: Art market; Primitive art; Natural language processing; Sentiment analysis; Taste studies
Topics: Global, Europe, English, 20th Century, database creation, management, and analysis, text mining and analysis, Art history

Art market studies have long relied on econometrics to explain the prices of works through a series of variables, such as the dimensions of the work, its medium, its date, whether or not its creator is alive, etc. However, this quantitative analysis fails to measure the qualitative opinion on the work, that can be found in exhibition or auction catalogues. As a consequence, no evidence is made between the price of an artwork and the critical comments on the latter, even if it seems obvious that such a link exists. This lightning talk aims at filling this gap by introducing Natural Language Processing, and more specifically, sentiment analysis, into econometrics (Généreux et al., 2008).

The study of opinions is all the more relevant as it relates to an emerging market: the paper will thus focus on the Parisian auction sales of so-called "modern" paintings in the 19th century - as opposed to the Old Masters market. A comprehensive transcription of all the related catalogues (for more details, see Saint-Raymond, 2018) allowed to apply sentiment analysis to all the descriptions of the works sold at auction. The minutes of the sales, curated at the Archives de Paris, then provided the hammer prices, which were added to this dataset. Finally, hedonic regressions were run, including sentiment analysis (Liu, 2018) as an explanatory variable, in addition to all the traditional ones. Sentiment analysis is based on a lexical analysis taking into account the specificities of our corpus.

This talk will be an opportunity to discuss our results and the relevance of this methodological encounter: how can one measure opinions on artworks and, more broadly, tastes?

References

Généreux, Michel, Poibeau, Thierry and Koppel, Moshe, 2018. "Sentiment analysis using automatically labelled financial news". LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology, Marrakech, Morocco.

Liu Bing (2015). Sentiment Analysis Mining Opinions, Sentiments, and emotions, CUP, Cambridge.

Saint-Raymond, Léa, 2018, "Le pari des enchères : le lancement de nouveaux marchés artistiques à Paris entre les années 1830 et 1939. Corpus bibliographique des ventes aux enchères publiques considérées", Bibliographic corpus of the considered public auctions, https://doi.org/10.7910/DVN/MZIBKB, Harvard Dataverse.

Acknowledgement

Thierry Poibeau is partially funded by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

What Is A Reader? How Readers on Goodreads are Changing the Canon in the Twenty-First Century

Ashley Champagne
ashley_champagne@brown.edu
Brown University, United States of America

Keywords: book culture, reading practices, machine intelligence
Topics: English, North America, Contemporary, crowdsourcing, social media analysis and methods, First nations and indigenous studies, Media studies

Matthew Kirshenbaum's essay "What Is An @uthor?" argues that today's social media landscape provides authors with a different means to confront their public personas.[1] Authors can tweet back to their readers, like William Gibson tweeted to an MLA panel in January of 2015 on his novel The Peripheral, to engage in immediate and mediated conversations about their work on a global scale. In this short paper I ask "What is a Reader?" because the Amazon-owned social media site, Goodreads, changes how readers respond to literature. While literary culture was once accessible primarily to those who had the means to acquire formal education, now literature is widely accessible and frequently free online; this shift in access is tied to an evolving version of literary knowledge in the digital age. There are free, online lectures about literature through TED Talks and Massive Open Online Courses (MOOCs), and discussion forums on the Internet are widely available through Goodreads, LibraryThing, Amazon, and Facebook's Year of Books. While "lowbrow" or popular literature is not easily separated from "highbrow" literature in any historical period, the difference between the two has never been so muddied as in the twenty-first century.[2] In this paper, I explore how the medium of Goodreads influences how readers respond to works by close and distant reading the comment section for two well-reviewed books on the site, Louise Erdrich's The Round House, a novel about the rape of a Native American woman by a non-Native man and Stephanie Myer's Twilight, a series about vampire romance. I situate the comments section for these two books within the broader context of comments posted on Goodreads about literature in order to highlight how online delivery systems can impact how and what we read. Because online delivery systems, like Goodreads, have the power to promote certain works over others, what constitutes literary value continues to change and therein their appeal to audiences beyond academic institutions. As a consequence, the taste of the large online public begins to define the classificatory matrix foundational to canon design; in turn, these seemingly "objective" classifications forged through algorithmic machine intelligence begin to fossilize a new canon. And yet, in circulating some texts selectively over others, these systems of counting and accounting define what counts as literature.

[1] Kirshenbaum, Matthew. "What Is an @uthor?" Los Angeles Review of Books. 6 February 2015. Accessed 7 Oct. 2019.

[2] One example of how "lowbrow" is not easily separated from "highbrow" literature in any historical period is the recent response to T.S. Eliot's The Waste Land. David Chinitz writes about Eliot's long-term interest in comic strips and crosswords (2003); Barry Faulk demonstrates how popular London music halls impacted Eliot's works (2011); and Robert Lehman writes on Eliot's use of satire in the drafts of The Waste Land (2009).

The Future of Storytelling in the Age of AI and Posthumanism

Serdar Tuncer
stuncer2@uottawa.ca
University of Ottawa, Canada

Boulou Ebanda De B'Beri
University of Ottawa, Canada

Keywords: Artificial Intelligence, Storytelling, Posthuman, Posthumanism, Anthropocene
Topics: Global, English, Contemporary, artificial intelligence and machine learning, natural language processing, Communication studies, Philosophy

Computer scientists are building a vast array of machine learning systems (often called Artificial Intelligence or AI) that can perform daily human tasks reliably, sometimes more so than humans. Indeed, AI is becoming capable of capturing the unique essence that makes us "human": storytelling. Humans are intrinsically

storytellers and able to express themselves through stories (Fisher, 1987; Frank, 2012). Moreover, people have always learnt about themselves and made sense of the rest of the world through narratives (Ebanda de B'béri, 2006; Tunçer, 2018). However, in recent years, AI-generated games and articles have already started to emerge (Fitch, 2019). While AI is able to exceed certain functions of natural human intelligence as it gains self-learning capacity in repetitive or analytical tasks, it may have the potential to harness the persuasive capabilities of storytelling.

What will become of humanity if AI can perform such "human" talents as storytelling? Indeed, some have gone as far as declaring that AI will be the greatest danger to humanity (Bramer, 2015). While other scholars have already hailed the arrival of the posthuman era (Wolfe, 2010; Braidotti, 2013) with such advances as neural nets that allow biological (human) and silicon-based artificial brain cells to communicate with each other (Serb et al., 2020). What will the future look like if AI can harness the influential potential of storytelling? What would be the story landscape in this era of posthumanism?

We argue these questions through the lens of persuasion and Transportation Narrative Theory, which refers to the experience of an individual being transported into a world created by a narrative (Green & Brock, 2000; Green, 2004). According to this theory, several mechanisms comprise the transportation experience: reduced counterarguing, connections with characters, heightened perceptions of realism, the formation of vivid mental imagery, and emotional engagement (Green & Fitzgerald, 2017). Through this transportation experience, a narrative can influence an individual's attitudes and beliefs. We argue that if or when AI develops the ability to create its own narratives, it will also have the potential to create transportation experiences and, thus, influence all aspects of human life.

REFERENCES:

Braidotti, R. (2013) The Posthuman (Cambridge, MA: Polity Press).

Bramer, M. (2015). AI Threats (Addressing Hawking and Musk). ITNOW, 57(2), 58-59.

Ebanda de B'béri, B. (2006). Mapping Alternative Expressions of Blackness in Cinema. Bayreuth, BASS/ Breitinger.

Fisher, W. R. (1987). Human communication as narration: Toward a philosophy of reason, value, and action. Columbia, S.C: University of South Carolina Press.

Fitch, A. (2019, October 13). Readers Beware: AI Has Learned to Create Fake News Stories. The Wall Street Journal.

Frank, A. W. (2012). Practicing dialogical narrative analysis. In J. A. Holstein, & J. F. Gubrium (Eds.), Varieties of narrative analysis (pp. 33-52). Thousand Oaks, CA: SAGE Publications Inc.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. Journal of Personality and Social Psychology, 79, 701-72.

Green, M. C. (2004). Transportation into narrative worlds: The role of prior knowledge and perceived realism. Discourse Processes,38, 247-266.

Green, M., & Fitzgerald, K. (2017, January 25). Transportation Theory Applied to Health and Risk Messaging. Oxford Research Encyclopedia of Communication. Retrieved 14 Jun. 2020, from https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-261.

Serb, A., Corna, A., George, R. et al. (2020). Memristive synapses connect brain and silicon spiking neurons. Sci Rep 10, 2590. https://doi.org/10.1038/s41598-020-58831-9

Tunçer, S. (2018). Reinterpreting Crisis Communications in The Post-Truth Era. Moment Journal of Cultural Studies (2018), 5(2): 368-382. https://doi.org/10.17572/mj2018.2.368382

Wolfe, C. (2010). What Is Posthumanism? Minneapolis: University of Minnesota Press.

Portugueses de Papel: Exploring Lusophobia and desidentificação in Brazillian literature

Nickoal Eichmann-Kalwara
Nickoal.Eichmann@colorado.edu
University of Colorado Boulder, United States of America

Pedro Cuevas-Collante
University of Colorado Boulder, United States of America

Kassandra McLean
University of Colorado Boulder, United States of America

Tania Martuscelli
University of Colorado Boulder, United States of America

Nivea Soto-Lightbourn
University of Colorado Boulder, United States of America

Carlos Martins-Filho
University of Colorado Boulder, United States of America

Keywords: Qualitative data analysis, Portuguese migration, Brazilian Literature
Topics: South America, English, 19th Century, 20th Century, bibliographic analysis, database creation, management, and analysis, Literary studies, Statistics

This presentation aims to share results from a data-driven approach for exploring Lusophobia and desidentificação (disidentification) in nineteenth- and twentieth- century Brazil. Combining qualitative data analysis and statistical approaches, we identified 169 Portuguese characters from a sample of 51 Brazilian books of the ~800 published during that time. In examining the characters and nuances of desidentificação, we interrogated 1) whether characters were represented positively neutrally, negatively or ironically; 2) whether their gender and occupation played a role in that representation, and 3) how portrayals of Portuguese migrants changed over time.

Contrasting Playwrights through Sonic Signatures

Oliver Calder
caldero@carleton.edu
Carleton College, United States of America

Eric Carlson Alexander
Carleton College, United States of America

Keywords: literature studies, linguistics, visualization
Topics: Europe, English, North America, 15th-17th Century, Contemporary, artificial intelligence and machine learning, natural language processing, Computer science, Literary studies

Well-written characters in drama and literature are often imbued with distinctive voices through differences in word choice, sentence structure, and even the sound of their words. In past work, we have shown that it is possible to detect phonetic differences between classes of Shakespeare characters (e.g., protagonists, antagonists) computationally, using Naive Bayes classifiers. In this talk, we will discuss our early attempt to use such classifiers to draw contrast between different playwrights. We lay out a corpus of early modern playwrights on an axis measuring the predictive accuracy of our model. Additionally, we present visual tools for identifying which phonemes are most distinctive for different characters and authors.

That's Not Vintage, It's Obsolete: Steve Reich's early e-sketches as a case study of current issues with early historical music notation software and why it matters

Twila Bakker
twila.bakker@gmail.com
Unaffiliated

Keywords: born-digital, manuscript, music
Topics: English, North America, 20th Century, Contemporary, media archaeology, music and sound digitization, encoding, and analysis, Library & information science, Musicology

Since the 1980s classical music composition has become increasingly digital. The working papers of the American composer Steve Reich, housed at the Paul Sacher Stiftung (Basel, Switzerland), demonstrate one such changing trajectory. Reich began his professional career in the 1960s with pencil and paper, and moved into a digitally assisted compositional practice in the 1980s. After Reich's adoption of the computer his working documents-in addition to handwritten sketchbooks and unbound sketches, audio tapes and assorted letters-become a blend of proprietary software files housed on 3.5" floppy disks and hand

marked up computer print outs. While some of these digital documents still function due to backwards compatibility, not all of the software features are still supported and the surviving software disks run only on vintage hardware. This paper explores some of the challenges and additional information that these born-digital files offer the contemporary musicologist.

Between Phallus and Freedom: An Ethnography on the Embodied Experiences of Tinder Users in Cape Town

Leah Davina Junck
leahjunck@gmail.com
University of Cape Town, South Africa

Keywords: Dating, Gender, Identity, Tinder
Topics: Africa, English, Contemporary, social media analysis and methods, Anthropology

This paper forms part of my PhD thesis, which I am currently in the process of finalising. It engages with dating applications the likes of Tinder and gender roles within this social space. The research context is Cape Town, a mutlicultural city in its current post-colonial, post-apartheid state, in which sexuality has only in recent years become a topic that may be openly discussed in certain social environments. In the face of current protests against gender based violence after a spade of brutal rapes and murders, it is very important to carefully and with look at the nuanced ways in which relationships are initiated via technologies and how technologically mediated encounters are perceived and engaged with.

Sober Hopes: Designing an Introductory Course for Computational Research in Hebrew / Arabic Texts

Itay Marienberg-Millikowsky
itay.marienberg@gmail.com
Ben-Gurion University of the Negev, Beer-Sheva, Israel

Keywords: digital humanities pedagogy, Hebrew and Arabic texts, distant reading, literary theory
Topics: Global, Europe, English, Contemporary, curricular and pedagogical development and analysis, meta-criticism (reflections on digital humanities and humanities computing), Education/ pedagogy, Literary studies

In this short presentation, I would like to demonstrate how a critical view of the Digital Humanities - or even an ambivalent or skeptical one - influenced the construction of a unique introductory course for computational text research, with an emphasis on literary texts. Thanks to a collaboration with the Open University of Israel, which undertook the publishing of the first introductory book in Hebrew, I had a rare opportunity to design a field of knowledge from its inception, in a particular cultural setting; the results, so I believe, would be relevant for anyone who wants to teach DH from a (traditional) humanistic standpoint, and to advance public digital humanities.

Sentiment sensitivity. Using sentiment analysis in literary studies to analyze genre and the depiction of illness

Evelyn Gius
Technische Universität Darmstadt, Germany

Anna Murawska
anna-aline.murawska@uni-hamburg.de
Universität Hamburg, Germany

Oliver Schmidt
Technische Universität Hamburg, Germany

Carla Sökefeld
carla.soekefeld@uni-hamburg.de
Universität Hamburg, Germany

Michael Vauth
Technische Universität Hamburg, Germany

Keywords: Sentiment Analysis, Character Analysis, Semantic Field
Topics: Europe, English, 19th Century, 20th Century, Contemporary, annotation structures, systems, and methods, text mining and analysis, Literary studies

Sentiment analysis for literary texts

Sentiment analysis has been used for fictional texts for about 15 years now, with a focus on plots and the development of characters. 1 Despite some interesting findings, sentiment analysis methods seem not to be directly deployable for literary text analysis yet. In addition to general problems (creation of suitable dictionaries, recognition of negations and other dependency-related changes in meaning), the high complexity of literary language and text structure as well as the considerably higher number of potential opinion targets or holders of sentiment (cf. Liu 2012: 18) are problematic.

In comparison to the still challenging high complexity of language (i.e., metaphors, irony, and other forms of improper speech), tackling text structure and sentiment holders seems more straightforward. A feasible use of sentiment analysis rooted in literary studies should therefore combine (i) simple, but established procedures for sentiment analysis with (ii) more elaborate understandings of texts and their entities.

We tried to implement this by developing the approach presented in this contribution.2 With two adjustments we made sentiment analysis more adequate for literary text analysis:
Segmentation: Since text segmentation should correspond to hermeneutic text approaches, we only analyze segments that are relevant for character analysis (for now: character speech) and combine them with thematic aspects. 3 Sentiment sensitivity: Due to the problems discussed, we conduct a comparatively simple sentiment analysis ignoring the positivity or negativity of sentiment values. We thus determine what we regard as sentiment sensitivity. Thereby, we avoid both attributing false sentiment values to negations or improper speech and positive and negative values cancelling each other out.

Implementation

For our two experiments we automatically extracted direct speech and manually annotated speakers for six exemplary texts from corpora of the hermA project4. We also created the semantic fields 'state and society' and 'illness'.5 Then a sentiment value was determined for each speech segment consisting of the sum of all values provided by the SentiWS sentiment lexicon (Remus et al. 2010), normalized by their number. Finally, speech segments including words from the pertaining semantic fields were identified.

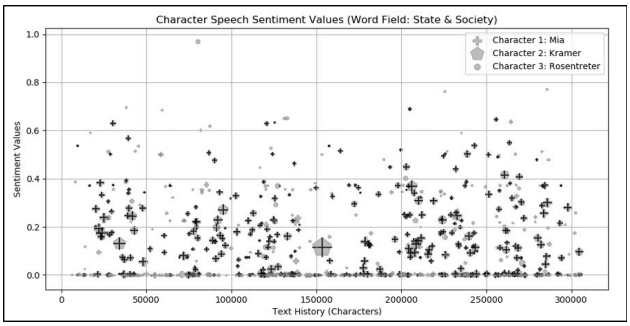


Figure 1: example of sentiment analysis in character speech based on semantic fields (Zeh: Corpus Delicti, crosses indicating semantic field segments)

Thus, we analyzed when, to what extent and with which sentiment values character speak as well as their relation to certain topics. The results concerning our research interests are sketched briefly below.

Example 1: Gender and Illness

In our first experiment with novels from around 1900, general sentiment sensitivity resulted higher for male than for female characters (cf. Table 1 and Figure 2). Interestingly, this difference almost levels out in segments containing the semantic field 'illness'. Therefore, while showing general differences between female and male speech around 1900, our sentiment approach proves useful for looking closer at the relation between gender and certain topics as well.

text	character group	character speeches	semantic field character speeches	Ø Token per Character Speeches	Ø Sentiment Value	Ø Semantic Field Sentiment Values
Aus guter Familie	female	463	126	18.4	0.098	0.156
	male	248	80	21.5	0.114	0.175
Das kleine Glück	female	582	143	21.23	0.078	0.139
	male	449	155	30.7	0.101	0.139
Ellernklipp	female	219	110	31.21	0.066	0.104
	male	323	159	34.5	0.094	0.146

Table 1: Overview of annotations and sentiment scores of Aus guter Familie, Das kleine Glückund Ellernklipp

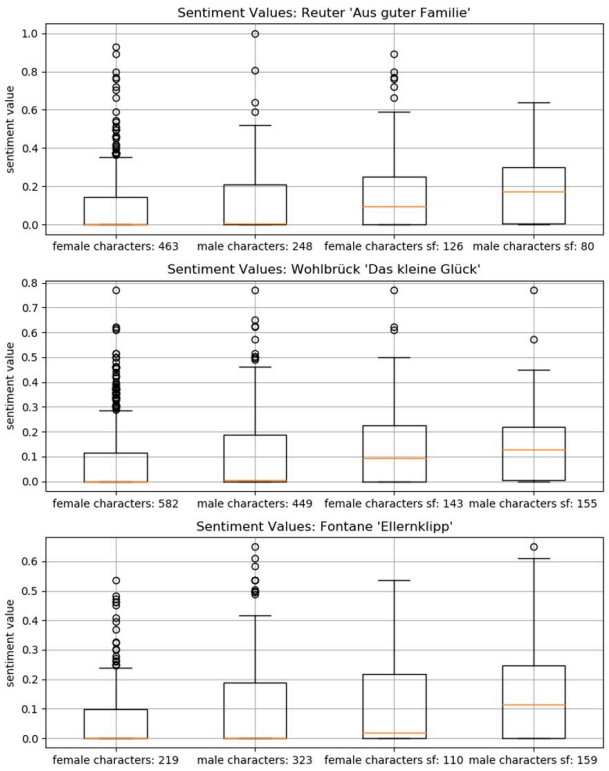


Figure 2: sentiment sensitivity of female and male characters from Aus guter Familie, Das kleine Glückund Ellernklipp(general and semantic field (sf)-related values)

Example 2 : Dystopias as Genre

In the second experiment we examined whether in contemporary dystopian novels the presumably crucial semantic field of state and society is sentiment sensitive. The values in Table 2 show that character speech about state and society in dystopias tends to be emotionalized. Thus, topic-related sentiment sensitivity seems to be an interesting feature for the automatic detection of dystopia as a genre.

Text	Character	Amount of speech segments	Ø sentiment value	amount of speech segments affiliated with semantic field[6]	Ø sentiment value (affiliated with semantic field)
Corpus Delicti	Mia	594	0.082	180	**0.140**
	Kramer	275	0.111	119	**0.158**
	Rosentreter	205	0.084	79	**0.133**
Die Optimierer	Samson	589	0.076	211	**0.126**
	Ercan	73	0.123	33	**0.152**
	Fleischmann	43	0.104	21	**0.151**
Black Out	Christopher	585	0.066	166	**0.112**
	Serenity	396	0.052	52	**0.087**
	Kyle	307	0.056	63	**0.101**

Table 2: sentiment sensitivity of character speech (with and without affiliation to the semantic field of 'state and society') in dystopian novels

In summary, our concept of sentiment sensibility, as calculated by sentiment analysis, seems to be an appropriate way of dealing with the ambiguity of literary texts and the interests of literary scholarship.

References

Primary Texts:

Elsberg, Marc: Black Out - Morgen ist es zu spät (2012). Fontane, Theodor: Ellernklipp (1881). Hannig, Theresa: Die Optimierer (2017). Reuter, Gabriele: Aus guter Familie (1895). Wohlbrueck, Olga: Das Kleine Glück (1910). Zeh, Juli: Corpus Delicti (2009).

Research Literature:

Adelmann, Benedikt; Andresen, Melanie; Begerow, Anke; Franken, Lina; Gius, Evelyn; Vauth, Michael (2019): Evaluation of a Semantic Field-Based Approach to Identifying Text Sections about Specific Topics. Digital Humanities 2019 Conference Paper.

Alm, Cecilia Ovesdotter; Sproat, Richard (2005): Emotional Sequencing and Development in Fairy Tales. In: Affective Computing and Intelligent Interaction: First International Conference, ACII 2005, Beijing, China, Proceedings, pp. 668-674.

Becker, Martin; Hettinger, Lena; Hotho, Andreas; Jannidis, Fotis; Reger, Isabella; Zehe, Albin (2016): Prediction of Happy Endings in German Novels based on Sentiment Information. In: Proceedings of DMNLP, Workshop at ECML/PKDD, pp. 9-16.

Bülthoff, Heinrich; Gerdemann, Dale; Meurers, Detmar; Mohler, Betty; Volkova, Ekaterina (2010): Emotional Perception of Fairy Tales: Achieving Agreement in Emotion Annotation of Text. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 98-106, 2010.

Burghart, Manuel; Dennerlein, Katrin; Schmidt, Thomas (2018): Kann man denn auch nicht lachend sehr ernsthaft sein? Zum Einsatz von Sentiment-Analyse Verfahren für die quantitative Analyse von Lessings Dramen. In: Book of Abstracts DHd 2018, pp. 244-248.

Jacobs, Arthur (2019): Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. In: Frontiers in Robotics and AI.

Landt, Matthias (2013): Sentiment Analysis as a Tool for Understanding Fiction (only available as pdf).

Liu, Bing (2012): Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies 16. San Rafael: Morgan & Claypool.

Mohammad, Saif (2011): From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural

Heritage, Social Sciences, and Humanities, pp. 105–114.

Remus, Robert; Quasthoff, Uwe, Heyer, Gerhard (2010): SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. In: Proceedings of the 7th International Language Resources and Evaluation (LREC'10), pp. 1168-1171.

[1] See for example the analysis of fairy tales (Alm 2005, Bülthoff 2010, Mohammed 2005), dramas (Burghart 2018), booklet novels (Becker et al. 2016), youth literature (Jacobs 2019) and fanfiction (Landt 2013).

[2] We have programmed all analyses in Python and will make them available with explanations and the resources used as Jupyter notebooks for subsequent use.

[3] In the next step of our work, we will also include character description.

[4] Cf. https://www.herma.uni-hamburg.de. Three texts were taken each from the corpora of the project "Gender and Illness" and the project "Dystopia as Genre" (cf. primary texts in the reference section).

[5] The semantic fields were created based on the approach presented in Adelmann et al. (2019).

[6] Speech segments containing at least one word from the semantic field are categorized as being affiliated to it.

Data-Driven In-class Activities for Chinese History Pedagogy

Lik Hang Tsui
lhtsui@cityu.edu.hk
City University of Hong Kong, Hong Kong S.A.R. (China)

Keywords: Chinese history pedagogy, digital humanities pedagogy, classroom activities, Chinese
Topics: Asia, English, BCE-4th Century, 5th-14th Century, 15th-17th Century, curricular and pedagogical development and analysis, Asian studies, Education/ pedagogy

My presentation focuses on the redesigning of classroom activities in undergraduate learning for Chinese history; it does so by examining and reflecting on the effectiveness of the in-class activities and assignments that I have devised and adopted for relevant courses. These include online annotation exercises of Classical Chinese texts with Hypothesis and CTEXT, extraction of biographical entries from the China Biographical Database, mapping exercises on WorldMap, and digital metrical analysis of classical poetry on the Souyun platform.

This study is funded by the University Grants Committee of Hong Kong's Teaching and Development Grants through a Teaching Start-Up Grant at the City University of Hong Kong: "Rediscovering Chinese History through Data: Incorporating Digital Humanities Pedagogy into Data-Driven Learning of Chinese History" (Project No.: 6000707).

Tracking the framing of politicians and news events across the multilingual Wikipedia translation and its unseen impact

Mark Shuttleworth
markshut@hkbu.edu.hk
Hong Kong Baptist University, Hong Kong S.A.R. (China)

Jun Pan
Hong Kong Baptist University, Hong Kong S.A.R. (China)

Zhilu Tu
Hong Kong Baptist University, Hong Kong S.A.R. (China)

Keywords: Wikipedia translation, multilingual Wikipedia, point of view, knowledge production, political translation
Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, social media analysis and methods, software development, systems, analysis and methods, Translation studies, Central/ Eastern European Studies

Major international news stories often give rise to the very rapid development of a multilingual set of

Wikipedia articles that are the site of multiple acts of translation, the presence of which can, however, be difficult to detect. Focusing on contemporary Russian news stories, this project has attempted to investigate this area, aiming to identify the presence of translation through the application of a number of different approaches and to interpret its function in terms of its contribution to knowledge creation and point of view. To date, the investigation has revealed much about the presence and role of translation in the encyclopaedia, while the next stages of the project are designed to extend, enrich and - possibly - challenge the existing analysis with the aid of a number of digital tools that we aim to develop.

A Prototype for a Truly Decentralized, Queryable LOD Store

Pascal Belouin
pbelouin@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science, Germany

Florian Kräutli
Max Planck Institute for the History of Science, Germany

Keywords: infrastructure, decentralized, linked open data, blockchain, archiving
Topics: Global, English, Contemporary, digital archiving, linked (open) data, Computer science, Library & information science

The sustainability, preservation, and long-term accessibility of research data is a growing concern for a number of academic institutions. Several centralized research data storage/hosting solutions are currently available to mitigate this issue, where research data archival and access is guaranteed by single centralized entities, often hosting research data for several academic institutions.

In contrast, a truly decentralized solution offers the benefits of removing the need to rely on a small number of centralized entities to store and access these data: this brings about several benefits such as censorship resistance, immutability of the data, the removal of a single point of failure, and the removal of a "direct" need to maintain the actual infrastructure necessary for long-term storage and availability.

In this context, the Max Planck Institute for the History of Science explores the development of a prototype that would permit the storage of research data in a truly decentralized way.

Deep Learning for Recognizing Chinese Historical Documents

Guangwei Zhang
zhangguangwei@snnu.edu.cn
Shaanxi Normal University, China, People's Republic of

Keywords: OCR, deep learning, pipeline
Topics: Asia, English, 5th-14th Century, artificial intelligence and machine learning, text mining and analysis, Cultural studies, Education/ pedagogy

Deep learning methods have greatly improved the text recognition accuracy for most modern languages. However, the OCR for historical documents is challenging, especially for the handwritten or printed ones without a training dataset, in which each character has hundreds or thousands of labeled examples. We focus on the recognition of the following types of primary sources for studying Chinese history, including handwritten Tangut historical documents, stele texts written in ancient Chinese, and place names on Chinese historical maps, as shown in Figure 1.

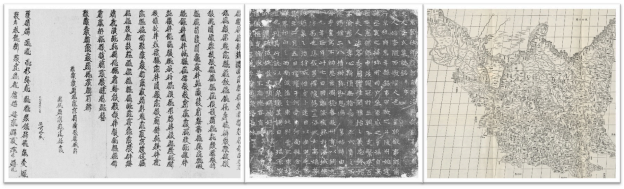


Figure 1. Tangut Historical Documents, Steles with Ancient Chinese Characters and Historical Maps

Tangut, invented and used by the West Xia State in Chinese history, has distinguished for more than 1000 years. The main Tangut documents were found and evacuated from Khara-Khoto by P. K. Kozlov in the early 20th century, which are important primary sources complementing to the Chinese historical documents for studying that period of Chinese history. A large collection of steles with texts in ancient Chinese keeps

important information of the time it was built, which needs to be digitized for history study. Historical maps are different from the first two types because they contain texts (place names) as well as images. The positions of place names on the map are random and the directions of the place names vary. Though the three tasks are different, we propose a unified workflow and framework for recognizing texts in these historical documents.

Our workflow has four phases: (1) text detection and segmentation, (2) character annotation, (3) model training, and (4) text line recognition. The core function of the workflow is deep convolutional neural networks (DCNNs). The multiply stacked convolutional layers are used in all the four phases of the workflow. They extract features that can be used in the character classification (phase 4) and text line detection (phase 1) as well as in the generation of supported characters for rare characters (phase 2). For the character classification and generation, a fully connected layer is added at the end of the stacked convolutional layers to output the predicted class for a given input image. For the text line detection, the fully connected layer is removed and a fully convolutional network (FCN) is attached to do pixel-level segmentation.

Text line detection and segmentation is a part of document layout analysis, where FCN [1]and U-Net[2] [3] are usually used. We use a modified U-Net to detect text lines in the Chinese historical documents and then segment them from the document images. The characters in the lines are annotated manually when we had not enough labeled single characters at first. We have developed a method to generate support examples for representing rarely used characters, Target-Directed Mixup for labeling characters[4]. When most characters have enough examples, we can synthesize a large number of text lines using the labeled characters to form the training dataset for the text line recognition. We train a model for recognizing the segmented lines by combining a convolutional neural network and a recurrent neural network based on the synthesized text lines with CTC[5] as the loss function. For the historical maps, the detected regions with places names are segmented and restored into normal orientation, and then they are fed into the trained model for recognition. The recognized text lines are shown in Figure 2 as an example, where human experts could correct the results.

Figure 2. Segmentation and Recognition Result of a Page of Tangut Documents

The workflow has been successfully used in these three tasks. Although the framework is designed for recognizing ancient Chinese and Tangut characters, which are similar in appearance, it could be modified and applied in recognizing other documents without an existing training dataset.

Acknowledgment

The author would like to thank the reviewers. This work is supported by MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No. 17YJCZH239).

References

[1] B. Barakat, A. Droby, M. Kassis, and J. El-Sana, "Text Line Segmentation for Challenging Handwritten Document Images using Fully Convolutional Network," in 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, Aug. 2018, pp. 374-379.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234-241.

[3] K. Ma, Z. Shu, X. Bai, J. Wang, and D. Samaras, "Docunet: document image unwarping via a stacked U-Net," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4700-4709.

[4] G. Zhang and Y. Zhao, "Target-Directed MixUp for Labeling Tangut Characters," in 2019 International Conference on Document Analysis and Recognition (ICDAR), Sep. 2019, pp. 202-207.

[5] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, Nov. 2017.

Co-occurrence of modal markers: a network analysis approach

Paola Marongiu
paola.marongiu@unil.ch
Université de Lausanne, Switzerland

Keywords: Latin, Modality, Diachronic Linguistics, Network Analysis
Topics: Global, English, BCE-4th Century, network analysis and graphs theory and application, Linguistics

This proposal issues from a project on the diachrony of modality in Latin. The object of interestis the co-occurrence of modal markers in a Latin corpus, with a focus on the syntactic structures orlinguistic elements with which they interact. I propose a quantitative and structural analysis ofthis phenomenon by means of network analysis techniques. Despite the applications in humanitiesand specifically in linguistics, few models have been designed in order to study targetedlinguistic elements in co-occurrence, and classical languages received even less attention.

The final goal of this proposal is to show how such intersections between computational methodsand theoretical linguistic matters would provide a new and deeper insight on both disciplines.

Making Uganda's Intellectual History Digital: Knowledge Preservation and Ethical Considerations

Samantha Emily Stevens-Hall
samantha.stevenshall@gmail.com
McMaster University, Canada

Keywords: Africa, intellectual history, OMEKA, digitization, ethical collaboration
Topics: Africa, Global, English, 19th Century, 20th Century, crowdsourcing, digital archiving, African and African American Studies, Book and print history

My DH project is an open access digital archive of primary sources and supplementary materials in African intellectual history.This archive serves as a repository for endangered documentary materials and as an exhibition to curate and display the intellectual history of Uganda. This archive will bring together scattered sources into one easily accessible online resource. For my dissertation digitizing pieces of intellectual history fosters discussion about the life cycle of documents, and what happens to sources when they are transferred from one medium to another. My thesis deals with themes of translation and the transition from oral to written culture, as well as what happens to the colonial archive and the dissemination of colonial knowledge when sources are made open access online. This project's mandate is to contribute to decolonizing the archive and bridging the "digital divide" between the West and Africa in computing access and capabilities.

Critical Making, Gender, and Textual Analysis: Beginning a Major Project

Amanda Henrichs
ahenrichs@amherst.edu
Amherst College, United States of America

Keywords: critical making, gender, text analysis
Topics: Europe, English, 3D printing, critical making, North America, 15th-17th Century, Contemporary, text mining and analysis, Design studies, Feminist studies

In this lightning talk, I will present images from a works-in-progress, specifically embroidered canvases that remediate information visualizations produced in the statistical program R. The information visualizations are based in a corpus of Early Modern text, namely works produced by the Sidney family (active 1580-1650). Two women in this family group, Mary Wroth and Mary Pembroke, were extraordinarily prolific and public writers in their historical and social contexts. I have already performed hierarchical cluster analysis on their texts in order to determine the intertextual relationship between the authors, and have embroidered several canvases that remediate the abstract computational product (the digital dendrogram) into a physical form (the embroidered object).[1]

This lightning talk presents the seeds of a project that integrates academic critical making, feminized crafting, and statistical textual analysis. Following Marie Hicks, Safiya Noble, Jacque Wernimont, and Marisa Parham, my working argument is that the embroidered canvases reveal that statistical natural language processing (traditionally associated with masculinized rationality) can be reframed as an explicitly feminist incursion into the larger digital humanities landscape.

I am interested in feedback from scholars and practitioners who work across multiple media forms, especially those with experience or interest in transitioning from traditional monographs toward more embodied and haptic forms of scholarship.

[1] For the original context of the dendrograms, see Amanda Henrichs, "Allusions in the Age of the Digital," Women Writers in Context https://wwp.northeastern.edu/context/#henrichs.allusions.xml

Eruptive Art Digital Analyses of Atmospheric Anomalies in American Painting

Melissa Schlecht
melissa.schlecht@ilw.uni-stuttgart.de
University of Stuttgart, Germany

Keywords: artworks, atmospheric anomalies, image analysis
Topics: Europe, English, North America, 19th Century, eco-criticism and environmental analysis, image processing and analysis, Art history

This paper focuses on the reconstruction of the geographical extension of atmospheric anomalies after volcanic eruptions across the United States as well as their influence on artistic productions during the 19th century. Methodically, I work at the intersection between classical art historical approaches and digital humanities, using computational tools that help analyze and evaluate the extensive image corpus of my study.

Subject Analysis of Cultural Heritage: Consuls in Taiwan and Diplomatic Incidents

Shu-hui Lin
siokhui@gapps.ntnu.edu.tw
National Taiwan Normal University, Taiwan

Keywords: Consul, THDL, Diplomatic, Cultural Heritage, 19th
Topics: Asia, Global, English, 19th Century, cultural analytics, digital archiving, Asian studies, History

Cultural heritage preserves the details of people's life around the globe. Consulates including the British Consulate at Takao and Fort San Domingo are all important cultural heritage of Taiwan. It is highly suggested that more exhibitions about consuls' life stories and their connections to Taiwan could be held in the future. The study collected material by searching information of consuls in Taiwan through Taiwan History Digital Library (THDL) and using DocuSky to build a theme lexical database. Enhancing the values of Taiwanese cultural heritage such as consulates helps promote Taiwan's experience in pursuing happiness to the world. It is hoped that the exploration of consuls in Taiwan, the establishment of consulates, and the relation between consuls and diplomatic incidents could facilitate the understanding of texts and the meanings, contexts of peripheral background information and further exert the influence of digital humanity on academic research and application.

Differential treatment how testimony is discussed in assault and sexual assault cases at the Court of Appeal of Québec

Louis Chartrand
lochartrand@gmail.com
University of Pittsburgh, Canada

Ulisce Desmarais-Grégoire
Desmarais Desvignes Crespo

Keywords: sexual assault, testimony, testimonial injustic, topic modelling, criminal law
Topics: English, North America, Contemporary, natural language processing, text mining and analysis, Humanities computing, Law and legal studies

Developments in the last decade have shed a new light on sexual crimes and raised the issue of police and justice response in the public consciousness. Much of the criticisms towards the justice system have focused on the testimony, which revictimizes the victim by forcing them to relive the events while having their credibility put in doubt. This suggests that sexual assault victims are treated very differently to victims of other crimes (Brody 2011). In this paper, we wish to substantiate, measure and qualify this claim in the context of Canadian criminal law through this question: is testimony being treated differently in sexual assault cases vs. in cases of assault which doesn't involve sexual activity? To address this question, we use topic models to identify discussions about testimony in court decisions from the Quebec Court of Appeal, and measure how they differ in sexual assault cases along various variables.

Leveraging Google Sheets and GitHub for Data Curation on the Princeton Ethiopian Miracles of Mary Project

Rebecca Sutton Koeser
rebecca.s.koeser@princeton.edu
Princeton University, United States of America

Nick Budak
nbudak@princeton.edu
Princeton University, United States of America

Rebecca Munson
rmunson@princeton.edu
Princeton University, United States of America

Keywords: data curation, data modeling, tool building
Topics: Global, English, Contemporary, data publishing projects, systems, and methods, physical & minimal computing, Humanities computing

The creation and curation of humanities datasets is an important scholarly activity that requires labor and expertise, and results in a research output that furthers scholarship (Elswit and Bench). As more scholars become interested in building datasets, we need better and simpler solutions for managing and publishing data. Many humanities datasets make sense as tabular or relational data, but not every scholar or project team has the skills, resources, or desire to create and manage a relational database.

As a possible solution to the gap between researchers' skills and technical requirements, we will demonstrate the tools we are prototyping to support data curation in The Princeton Ethiopian Miracles of Mary Project, adding lightweight infrastructure around Google Sheets and GitHub using generalizable scripts. Our approach is to model and structure the data as if implementing it in a relational database, but with the goal of creating a set of related sheets in a single Google Sheets spreadsheet with data validation to link them and avoid redundant data entry (Belcher et al.). We will show a Google Apps Script project that can create a new spreadsheet with configured sheets, fields, and data validation based on a JSON data structure. We will also demo a script that generates a regular, automatic export of the Google Sheets data and commits it to a GitHub repository, resulting a versioned copy available for querying, visualization, automated validation, interface prototyping and publication, leveraging static site technology and minimal computing principles, and eventual data deposit for publication.

Works Cited

Belcher, Wendy Laura, Rebecca Sutton Koeser, Rebecca Munson, Gissoo Doroudian, and Meredith Martin. "CDH Project Charter - Princeton Ethiopian Miracles of Mary 2019-20." Center for Digital Humanities at Princeton, August 2, 2019. https://doi.org/10.5281/zenodo.3359178.

Elswit, Kate, and Harmony Bench. "Datasets Are Research." Dunham's Data, September 5, 2019. https://www.dunhamsdata.org/blog/datasets-are-research.

Princeton-CDH/pemm-scripts. Python. Center for Digital Humanities at Princeton, 2020. https://github.com/Princeton-CDH/pemm-scripts.

Princeton-CDH/pemm-data. Center for Digital Humanities at Princeton, 2020. https://github.com/Princeton-CDH/pemm-data.

Evaluating Semantic Relation Alignment in Historical Bibles using Historical Embeddings

Maria Moritz
maria.berger@dfki.de
German Research Center for Artificial Intelligence

Keywords: semantic relations, historical embeddings, alignment
Topics: Europe, English, 15th-17th Century, 18th Century, 19th Century, natural language processing, semantic analysis, History, Informatics

Current techniques for typical NLP tasks such as plagiarism detection and paraphrase prediction do not necessarily work well for historical language text due to the heavy modification (caused by cul- tural and linguistic change, induction of errors, se- mantic change in meaning) that historical text is encountered with. To perform reuse (e.g., cita- tions, allusion) detection using machine or deep learning techniques, models are rare and text for a specific time period of genre exists seldom.

Previously, we modeled changes in parallel, historical English Bibles determining the ratio and type of modification. This work gives an under- standing of the degree, type and partition of mod- ification in historical text, which again is impor- tant to consider for example upfront a stylometry task. Modification can reach from inflection to replacement of semantically related words. Our approach therefore used normalizing tools and looked-up lexical databases (e.g., WordNet and BabelNet Navigli and Ponzetto (2010)) to identity semantic relations (synonymy, hyper(o)nymy, co- hyponymy, see Fig. 1). We found 90% of the rela- tions that two statistically aligned words share.

Now, we attempt to relate the last 10% using embedding representations of words from one pre-existing embedding model-based on historical English ranging from 1800 to 1900, henceforth StanHistEmb(1)- and one self-trained model- based on the Royal Society Corpus (Kermes et al., 2016)(2) ranging from 1665 to 1869, henceforth RoySocEmb). We think the latter performs bet- ter being tailored well to the actual historical vo- cabulary. We complete the work with first manual evaluation.

Before, Faruqui et al. (2015) refined word vec- tor representations using precise relations of semantic lexicons. Evaluation against shows a per- formance increase of 13% compared to earlier ap- proaches. Naya (2015) use neural models to pre- dict the correct hypernym of a word in a vec- tor space and achieves an accuracy of 20% for predicting whether the correct vector is among the 10 nearest. Weeds et al. (2014) train an SVM on word vector pairs achieving an error-rate-reduction of about 15% and find that distance scores between paired vectors work well for en- tailment tasks while adding vectors is useful for co-hyponymy determination.

Data Used: We use aligned Bibles from 1500-1900 (each with each other). Berkeley Aligner (Liang et al., 2006) aligns verse-aligned Bibles on token-level (see Figure 1).

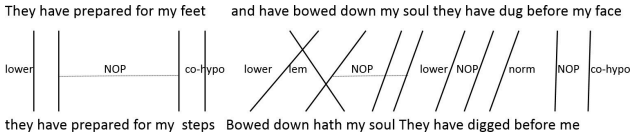


Figure 1: Modification examples during alignment

Recall Using StanHistEmb: First, we calcu- late the cosine distance for two word vectors us- ing the StanHistEmb vectors attempting to retrieve both words from each couple in the model. We find a cosine distance to 260,641 couples, which are normally distributed over the distance scale (see Figure 2, curve of o).(3)

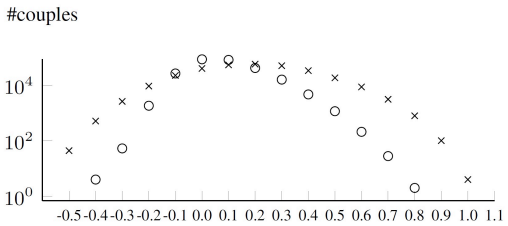


Figure 2: Distribution of cosine distance measures of the word couples determined with the StanHistEmb (o) and the RoySocEmb (x) model

Recall Using RoySocEmb: Now, we calcu- late the cosine similarity for two words using a self-trained RoySocEmb model and the Gensim li- brary. We find a cosine score to 302,812, much more than using StanHistEmb. Here, couples are even better normally distributed (Figure 2, curve of x), but values are lower due to a better tailored model, which in the sum is much smaller while supporting more historical vocabulary.(4)

Precision Using StanHistEmb: We manually evaluate 138 sample couples that have a cosine similarity(5) > .5. We assign 1.0 when two words are synonyms and 0.0 when there is no semantic relation(6) (see Table 1).Principally, scores from 0.0 to 0.4 represent false positives. While many couples scored above .5 are strongly related or synonym. We not yet manually evaluated sam- ples for RoySocEmb. However, we looked up the same samples using that model and find 119 out of which 69 scored > .5 and 77 scored > .4. How- ever, looking at the skipped couples, they tend to be false positives for high similarity.(7)

manual score	no. of couples	example
0.0	0	-
0.1	1	thyself,wilt
0.2	3	hast,thine
0.3	0	-
0.4	0	-
0.5	1	firebrands,darts
0.6	2	forgive,pray
0.7	27	wife,daughter
0.8	70	openly,publicly
0.9	34	hip,thigh
1.0	0	-

Table 1: Semantic relation score between word couples filtered using a cosine similarity of > .5 using the StanHistEmb model

Forecast: We plan to investigate the selftrained model of Early Modern English more deeply. Specially to see in more detail whether and how the size of the model influences the accu- racy of the words' relations. Therefore, we con- sider to refine the pre-trained model by retrain- ing and consider joining both models, because the RoySocEmb fits our time, but not our domain.(8)

References

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1606-1615.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The royal society corpus: From uncharted data to corpus. In LREC.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In Proceedings of the Human Language Technology Conference of the NAACL, Main Conference, pages 104-111. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual seman- tic network. In Proceedings of the 48th annual meet- ing of the association for computational linguistics, pages 216-225. Association for Computational Lin- guistics.

Neha Naya. 2015. Learning hypernymy over word em- beddings. In Deep Learning for Natural Language Processing - Project Reports. Stanford University.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hyper- nyms and co-hyponyms. In Proceedings of COL- ING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 2249-2259. Dublin City University and Association for Computational Linguistics.

1) https://nlp.stanford.edu/projects/histwords/

2) Containing text from articles, reviews and abstracts

3) 351,169 contain out-of-vocab words; 64,305 are as- signed "nan" through by-zero-division

4) 373,303 couples were not retrieved using RoySocEmb.

5) We normalize the distance score.

6) Evaluation is done by a German native with proficient knowledge in English.

7) We consider to fold the scores in future experiments down to a scale of only three or four digits. For now, we follow the value range between zero and one.

8) In accordance with the RSC, we plan to publish our embedding models.

An Online Odyssey: Digital Storytelling in the Humanities Classroom

Laura Biesiadecki
tuf84653@temple.edu
Temple University, United States of America

Keywords: Digital Storytelling, Pedagogy, Literature
Topics: Comparative (2 or more geographical areas), English, 20th Century, Contemporary, curricular and pedagogical development and analysis, electronic literature production and analysis, Literacy, composition, and creative writing, Literary studies

In the spirit of "carrefours/intersections," I propose a short presentation highlighting the instructional

benefits of interactive digital content in the humanities classroom. The primary source for this session will be a paper focused on addressing pedagogical concerns through digital methods, and detailing the process of creating visually dynamic and accessible content for literature classes. To supplement the paper, I will present a website designed to engage teachers and students in undergraduate instruction of James Joyce's "Ulysses." As a work in progress, the website will be published to showcase potential uses of different design techniques and data visualization tools.

Behind the Scenes of the Spectators Uncovering Anonymous Authors in Periodicals of the Enlightenment

Elisabeth Hobisch
elisabeth.hobisch@uni-graz.at
University of Graz, Austria

Martina Scholger
martina.scholger@uni-graz.at
University of Graz, Austria

Alexandra Fuchs
University of Graz, Austria

Bernhard C. Geiger
Know-Center Graz, Austria

Philipp Koncar
Graz University of Technology, Austria

Sanja Saric
University of Graz, Austria

Keywords: Spectators, digital edtion, stylometry, anonymous authorship, 18th century
Topics: Europe, English, 18th Century, attribution studies and stylometric analysis, scholarly editing and editions development, analysis, and methods, Cultural studies, Literary studies

The Spectator press is a pan-European periodical genre of the 18th century renowned for the deliberate play of authorial disguise: instead of writing under their real names, the spectatorial authors introduce a fictitious author-character. In the most prominent model of these periodicals - The Spectator (1711-1712, 1714) - the real authors, Joseph Addison and Richard Steele, create the fictitious author-character of 'Mr. Spectator' who, as an anonymous and confidence inspiring observer, communicated with the readers by exchanging letters with them. Yet, not only Addison and Steele but also other collaborators hid behind the eidolon of 'Mr. Spectator'. However, the real authors tagged the individual issues by a code of letters, which enables the readers of the time and contemporary literary scholars to attribute issues of respective periodicals to a specific author (cf. Rau 1980, 74).Throughout the 18th century, all Spectator emulations adopted the mask of the fictitious author, but largely lacked these authorial markers, making it difficult to attribute them to a specific author with close reading literary analysis methods. To counteract this problem, distant reading methods, such as stylometry, can provide significant insights into the text corpus of the pan-European Spectator press, especially regarding the attribution of presumably anonymous and collaborative authorship.The publicly available digital edition of the European Spectators, implemented at the University of Graz, contains more than 3.500 single issues of periodicals in Spanish (647 issues), French (1672 issues), and Italian (1320 issues) which are encoded using the standard of the Text Encoding Initiative. Many of the periodicals are published anonymously or by a group of authors. Based on this corpus, the Distant Spectators project, a cooperation of the Centre for Information Modelling, Institute of Romance Studies (both University of Graz), the Institute of Interactive Systems and Data Science (Graz University of Technology) and the Know-Center Graz, examines the application of various distant reading methods to a multilingual literary corpus and cross-references the results with a well established close reading literary research expertise. Our methods and results developed in the course of this project will also be open sourced, enabling other researchers in the field to conduct similar studies.In this contribution, we discuss how the application of specific stylometric methods can be employed to infer the identities of anonymous authors or translators, while bearing in mind the limitations of only a few comparable texts of the individual authors and the relative brevity of single issues of the periodicals. Specifically, we start by

computing Burrow's Delta as a first baseline approach. Our features include tuples consisting of words and their respective part-of-speech tags. We then use all issues of Spectator periodicals for which the authorship is clearly known and create hold-out test sets for each of them (depending on the number of issues written by respective authors with a maximum test set size of 10 issues per author). Our results are promising, achieving a mean performance of 81% across all three languages, meaning that, on average, we can detect the correct author for 81% of issues written in a respective language. This suggests that such methods are applicable to the Spectator periodicals. An outstanding example of unsolved anonymous redaction is the French translation of The Spectator, Le Spectateur ou le Socrate moderne, which was published anonymously between 1714 and 1726. While Martens (1968, 25) suggested a possible participation of Van Effen in the translation of the text, Rau (1980, 153-159) and Lévrier (2007, 42-44) considered it unlikely. To investigate this case through stylometric methods, we first compute similarities between texts using the Stylometry with R package based on term frequencies of all French periodicals from known authors and Le Spectateur ou le Socrate moderne, representing the result as a bootstrap consensus network (Eder 2017). Here, the translations are close to works of Van Effen and La Chapelle (cf. Figure 1), partly confirming the theory of Martens (1968, 25).

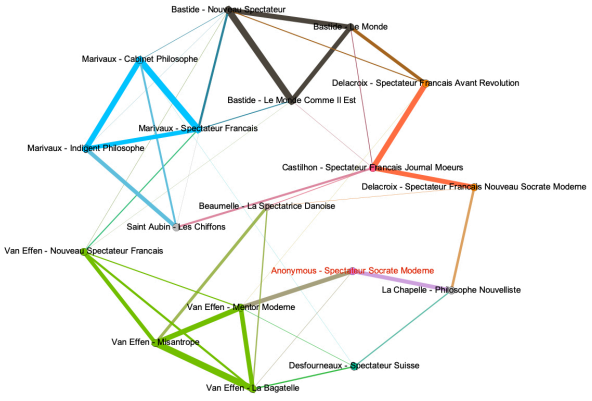


Figure 1: Bootstrap Consensus Network of French Spectator periodicals

Due to the fact that several authors published translations as well as their own writings, in this analysis we represent distinct periodicals of all authors as separate nodes. Note that the grouping of all three publications of Van Effen, but also those of Bastide, Marivaux and Delacroix indicate that the individual style of the authors is detectable through the network. Also, we note the close clustering of Le Mentor Moderne (translation of The Guardian by Van Effen), Le Philosophe nouvelliste (translation of The Tatler by Boisbeleau La Chapelle) and the anonymously published Le Spectateur ou le Socrate moderne. This observation could indicate that the style and choice of vocabulary in this example was determined by the style of the original texts, all originating from the English authors Joseph Addison and Richard Steele, as suggested by Rybicki (2012).To verify these results and further improve authorship detection we employ the General Imposters method comparing the Le Spectateur ou le Socrate moderne with texts from possible authors. Using different distance measures and most frequent words samples revealed Van Effen as the most likely author with certainty values between 0.82 and 0.95 which corroborates Martens claim. Our contribution showcases the combination of distant and close reading approaches to uncovering the authorial disguise and authorship attribution in periodicals of the Enlightenment. It can serve as a framework for cross cultural analysis, helping to better understand the history and spreading of ideas emerging during the Age of Enlightenment and their transnational diffusion in Europe.

Bibliography:

Addison, Joseph/Richard Steele (1711-1714): The Spectator. Edition by Donald F. Bond(1965), 5 volumes, Oxford: University Press.

Anonym (1716-1726): Le Spectateur, ou le Socrate moderne, Où l'on voit un Portrait naïf des Mœurs de ce Siècle. Traduit de L'Anglois. Paris: Etienne Papillon/ Frères Wetstein/François-Guillaume L'Hermite. Edition by Ertler/Fischer (2011-2019).

Eder, M. (2018). Authorship verification with the package stylo. https://computationalstylistics.github.io/docs/imposters

Eder, M. (2017). "Visualization in stylometry: cluster analysis using networks." Digital Scholarship in the Humanities, 32(1): 50-64

Eder, M., Rybicki, J. and Kestemont, M. (2016). "Stylometry with R: a package for computational text analysis." R Journal, 8(1): 107-21. https://journal.r-project.org/archive/2016/RJ-2016-007/index.html

Ertler, K.; Fuchs, A.; Fischer, M.; Hobisch, E.; Scholger, M.; Völkl, Y. (eds.) (2011-2019): The Spectators in the International Context, https://gams.uni-graz.at/spectators.

Lévrier, A. (2007): Les journaux de Marivaux et le monde des « spectateurs ».Paris: Presses de l'Université de Paris-Sorbonne.

Martens, W. (1968): Die Botschaft der Tugend. Die Aufklärung im Spiegelder deutschen Moralischen Wochenschriften. Stuttgart: Metzler.

Rau, Fritz (1980): Zur Verbreitung und Nachahmung des 'Tatler' und 'Spectator'. Heidelberg: Winter.

Rybicki, Jan (2012): "The great mystery of the (almost) invisible translator. Stylometry in translation", Quantitative Methods in Corpus-Based Translation Studies, edited by Michael P. Oakes and Meng Ji. John Benjamins Publishing Co, 231-248.

The dynamics of Czechoslovakia's policy as a reaction to the influence of the USSR and the Warsaw Pact countries on the materials of the archive of special services and the periodical press of Czechoslovakia 1968-1989.

Anatoly Vladimirovich Iashchenko
yaschenko.anatoliy.ay@gmail.com
Perm State University, Russian Federation

Keywords: Historical politics, computer source studies, stylometry, GIS, Czechoslovakia.
Topics: Europe, English, 20th Century, attribution studies and stylometric analysis, data modeling, History, Central/Eastern European Studies

Relations between Czechoslovakia and the USSR are an important block in the political history and history of international relations. The study of this problem field allows us to rethink the ideological conflict that arose during the relationship between the two states, the echoes of which are felt in the modern world in the form of vandal actions directed at monuments, at the joint historical and cultural heritage of the USSR and Czechoslovakia, to once again urge the public to pay attention to existing problems originating in the past.

The study uses the methods and approaches of historical informatics, computer source studies, and computer linguistics. This methodological synthesis allows you to take a fresh look at archival sources and engage in the analysis of previously unexplored information potential of documents and periodicals. And to create, based on the new data, dynamic models for constructing politics in Czechoslovakia.

Book printing and the rise of the vernacular in Europe, 1500–1820

Jani Marjanen
University of Helsinki, Finland

Hege Roivainen
University of Helsinki, Finland

Leo Lahti
Univeristy of Turku, Finland

Mikko Tolonen
mikko.tolonen@helsinki.fi
University of Helsinki, Finland

Keywords: book history, bibliographies, vernacularization, nation building
Topics: Europe, English, 15th-17th Century, 18th Century, 19th Century, bibliographic analysis, cultural analytics, Book and print history, History

Throughout Western Europe, the dominance of Latin as the language of written culture eroded and local languages became carrieres of politics, religion, culture and science in the early modern and modern periods. This process is sometimes called vernacularization, but it was not a unified process. We aim to describe this process in a more precise way by using library catalogues as data for language use in Europe from the sixteenth century to the early nineteenth century. We further zoom in on different types of cities in different language areas and discuss different publication profiles of university towns, capital cities and commercial centers.

Corpus-DB: a Scriptable Textual Corpus Database for Cultural Analytics

Jonathan Reeve
jonathan.reeve@columbia.edu
Columbia University, United States of America

Keywords: corpora, corpus, APIs, database, textual corpora
Topics: Comparative (2 or more geographical areas), Global, English, North America, 19th Century, 20th Century, Contemporary, information retrieval and querying algorithms and methods, text mining and analysis, Humanities computing, Literary studies

Corpus-DB (corpus-db.org) is a database and query framework which solves the problems of text retrieval, text cleaning, corpus compilation, and metadata aggregation that often form the first step for researchers in computational text analysis. Traditionally, scholars interested in studying a collection of texts, such as: novels set in London, Bildungsromane, sestinas, or poems written in 1889, have had to manually assemble their corpora, which can be a prohibitively laborious process. Corpus-DB gathers full texts and metadata from Project Gutenberg, the British Library, and other sources; cleans the texts; adds metadata found via Wikidata, Goodreads, and Wikipedia, and elsewhere; and provides this as a free, open, and easily scriptable API. This enables rapid prototyping of text analysis projects, as well as advanced queries of these corpora, providing easy answers to questions such as the average Goodreads star rating of novels set in London, or the median publication date for detective novels.

"Network Analysis and Spatial Stylometry in American Drama Studies" (NASSA)

Dennis Mischke
dennis.mischke@uni-potsdam.de
University of Potsdam, Germany

Michał Choiński
michalchoinski@gmail.com
University of Potsdam, Germany

Keywords: Network Analysis, Spatial Stylometry, GIS Mapping, Digital Drama Analysis
Topics: Europe, English, North America, 18th Century, 19th Century, 20th Century, cultural analytics, text mining and analysis, Literary studies, Performance Studies: Dance, Theatre

Dr. Dennis Mischke, University of Potsdam (Potsdam, Germany), Dr. Michał Choiński Jagiellonian University (Kraków, Poland), Joanna Byszuk, Institute of Polish Language, Polish Academy of Sciences (Kraków, Poland), Mathias Göbel, Staats- und Universitätsbibliothek (Göttingen, Germany)

Abstract:

"Network Analysis and Spatial Stylometry in American Drama Studies" (NASSA)

The transnational workgroup "NASSA" (NASSA, 2019) seeks to promote computational research methodologies in American Drama Studies. In particular, the workgroup develops a digital workflow that combines stylometry, network analysis and GIS mapping to conduct a comprehensive digital study of the Chadwyck-Healey American-Drama Corpus 1714-1915 (ADC-CHC 2006) - a collection of 1500 American dramatic texts from three centuries. The project aims at the accomplishment of three overall research goals. First (A), we want to account for a number of about 150 texts by anonymous authors by measuring their stylometric distance to texts by known playwrights of the same corpus, using the Burrows' Delta method (Burrows, 2002). Second (B), we want to study potential correlations between the linguistic origin (such as regional dialects) of authors (cf. Choiński, Eder & Rybicki, 2018) and the non-fictional spatial settings of a selected number of dramatic texts in a method that we call spatial stylometry. We propose spatial stylometry as a cross-section between geospatial humanities and stylometry. In our approach we combine the automatic extraction of geonames via Named-Entity Recognition (Spacy, NLTK) and match these entities with open repositories of geonames (such as the OpenStreetMap database). In a further step, we finally visualize the annotated results on a map that displays detected stylometric signals related to the spatial positioning of particular dramas. With this workflow, we wish to inquire whether detected real-word settings of a selected number of dramatic texts cluster in meaningful ways, approaching questions such as: are texts by playwrights from distinct

regions also set in these places? and how do detected similarities between texts correspond to authorial and metatextual information such as race, class or gender? In a third avenue (C), we seek to enhance our study with a network analysis and distant reading of scenic presence of characters (cf. Trilcke, Fischer 2018) to find potential correlations between the social configuration of individual texts, their spatial styles as well as racial, ethinic or gender-specific configurations.

References:

ADC-CHC: American-Drama Corpus 1714-1915. Chadwyck-Healey Collection ProQuest, URL: http://collections.chadwyck.com/marketing/products/about_ilc.jsp?collection=amdram (2006)

Burrows, J. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing, 17(3), (2002), 267-287.

Choiński, Michał, Rybicki Jan, Jonathan Edwards and Thomas Foxcroft. Pursuing Stylometric Traces of the Editor". Amerikastudien. No. 63 (vol. 2), pp. 141-158. URL: https://amst.winter-verlag.de/data/article/8299/pdf/101802005.pdf

Choiński, Michał, Eder Maciej & Rybicki Jan, Computational Criticism and Contemporary American Southern Fiction. Paper delivered at European Association of American Studies Conference, King's College (London), 4-7 April 2018.

Eder, M., Rybicki, J., Kestemont, M. Stylometry with R: A Package for Computational Text Analysis. The R Journal 8, (2016): 107-121.

Heuser, R., Moretti,F., Steiner, E., "Mapping London's Emotions." New Left Review 101 (2016): 63-91.

NASSA: Network Analysis and Spatial Stylometry in American Drama Studies. URS: https://www.uni-potsdam.de/de/digital-humanities/aktivitaeten/projekte/network-analysis-and-spatial-stylometry-in-american-studies-nassa.html (2019)

Trilcke, Peer. Fischer, Frank. Literaturwissenschaft als Hackathon. Zur Praxeologie der digitalen Literaturwissenschaft und zu ihren epistemischen Dingen, in: Sybille Krämer, Martin Huber (Hg.): Wie Digitalität die Geisteswissenschaften verändert: Neue Forschungsgegenstände und Methoden (= Sonderband der Zeitschrift für digitale Geisteswissenschaft 3), 27.6.2018, DOI: http://dx.doi.org/10.17175/sb003_003

The Dutch Veterans Memoires Work Bench: Bridging the gap between a traditional and computational reading of memoirs

Stefania Scagliola
stefania.scagliola@uni.lu
University of Luxembourg

Peter Verhaar
University of Leiden

Keywords: military memoirs, text mining, digital pedagogy, mix methods
Topics: Asia, Europe, English, 20th Century, curricular and pedagogical development and analysis, text mining and analysis, Education/ pedagogy, History

Dutch veterans have contributed to a broad collection of published memoirs on the topic of the independence war with Indonesia between 1945 and 1949. The Centre for Digital Scholarship has processed the corpus of 650 memoirs into machine readable text via OCR, and created a series of python scripts in a Jupyter Notebook environment. As two researchers, one specialised in text mining, and the other in military and digital history, we want to explore how we can encourage traditional historians to engage with these sources with text mining tools via a series of workshops. in our talk we inted to expand on the feasibility of developing a model for the integration of traditional and computational approaches to analyzing military memoirs. This could be of added value for the integraton of digital methods in the traditional history curriculum and research trajectory.

TexTiles: Visualizing the Patterns of Discourse

Robert Roessler
robertroessler@g.harvard.edu

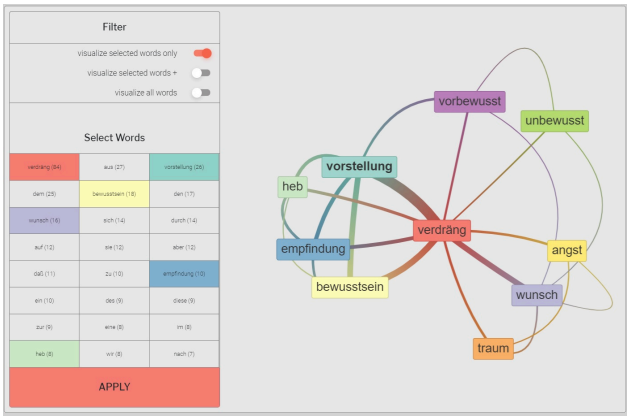
Harvard University, United States of America

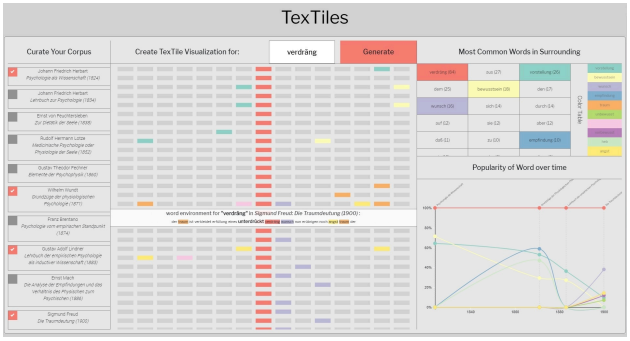
Michael Behrisch
Harvard University, United States of America

Johanna Beyer
Harvard University, United States of America

Keywords: Digital Discourse Analysis, Data Visualization in the Humanities, Keywords in Context Analysis
Topics: Europe, English, 19th Century, Interface design, development, and analysis, text mining and analysis, History of science, Literary studies

Discourse analysis is a well-established method in the humanities to analyze historical trends and knowledge figurations that are inscribed in the texts of an era. While these trends are manifested as linguistic nuances in a variety of ways, the actual discourse remains a rather abstract concept. The goal of our paper is to develop a visual representation of such a discourse. We present TexTiles, a Visual Analytics framework that allows domain specialists to visually trace historical discourses in large text corpora. The prototype allows scholars to curate a customizable corpus, analyze keywords in context, and explore the network of these context words within the corpus. Allowing for a hybrid approach between close and distant reading practices, we demonstrate the utility of our application based on a case study on the discovery of the unconscious and the mechanics of repression in the long 19th century.





Victimisation and otherisation in Cameroon Anglophone literature: A Study of Selected Authors from the 1980s to Present

David Toh Kusi
kusitoh@gmail.com
The University of Bamenda, Cameroon

Keywords: Diatribes, victimisation, otherisation, anglophone, literature
Topics: Africa, English, 20th Century, Contemporary, cultural analytics, text mining and analysis, Humanities computing, Literary studies

Literary works notoriously reflect the space and time of their production. In the case of Cameroon, the colonial legacy that set a divide between English Speaking and French Speaking communities has fostered two distinct, sometimes conflicting

identities. The paper, therefore, attempts a sentiment analysis of Cameroon Anglophone discontent through victimisation and otherisation in four plays and three novels distributed over 40 years of post-independence.

Creating New Music with Big Data and Evolutionary Algorithms

Jeffrey A T Lupker
jlupker@uwo.ca
The University of Western Ontario

William J Turkel
The University of Western Ontario

Keywords: Music; Big Data; Artificial Intelligence; Musical Composition; Music Theory
Topics: Europe, English, North America, Contemporary, artificial intelligence and machine learning, music and sound digitization, encoding, and analysis, Humanities computing, Musicology

Creating New Music with Big Data and Evolutionary Algorithms

Jeffrey A. T. Lupker & Dr. William J. Turkel

The availability of large musical datasets like the Million Song Dataset (Bertin-Mahieux et al 2011) has made it possible for researchers in Music Information Retrieval (MIR) to develop artificial intelligence (AI) techniques that perform a wide variety of musically-based tasks. In part, this is because algorithmically-determined characteristics of a song, including its key, mode, pitch, tempo and timbre, can be used to create labelled instances for training supervised learners like various kinds of neural nets. At the same time, the existence of these large datasets has made it possible for researchers to take a 'big data' approach to various styles of western music. One notable example is the work by Serra? et al (2012) which showed the changes and trends related to pitch transitions, the homogenization of the timbral palette and growing loudness levels that have shaped pop music over the past 60 years. The authors went on to suggest that past songs might be modernized into new hits by restricting pitch transitions, reducing timbral variety and making them louder.

We believe that it is possible to make use of musical big data to develop new kinds of compositional tools that employ AI to aid in the creation of new music. Although there is certainly a tradition of using computers in composition, many tools that have been created without a grounding in musical theory have often failed to attract the attention of composers. Existing examples of AI that have been implemented in music composition have largely been directed toward autonomously generated music and not toward tools that might aid a composer (Briot et al., 2019).

Here we present work-in-progress that attempts to bridge the gap between the big data approach, which is based on algorithmically determined physical properties of a song, and an approach grounded in strong musical theory. From the perspective of the music theorist and composer, a number of techniques that are common in the MIR literature do not always make much musical sense. We have been able to improve performance of mode and key detection algorithms, for example, simply by incorporating more musical knowledge into them. Our long-term goal is to determine how a series of AI algorithms trained to learn from pitch and timbre relationships according to mood, might aid and benefit a composer in the process of creating new music. What can the study of pitch transitions and timbral relationships offer the creation of new mood-based music compositions?

The system that we will discuss uses a genetic algorithm trained on pitch and timbre relationships from a large dataset. Given a desired mood, the program makes suggestions on musical structuring, chord progressions, timbre options based on chord selections and even melodic segment suggestions given in real-time, akin to the "autocorrect" feature on smartphones. As each suggestion is offered by the program, the composer can select whether or not they like it, dislike it or feel neutral towards it, so the program can continue to learn from the composer with each iteration of the genetic algorithm.

References

Bertin-Mahieux, Thierry, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset." In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011. http://millionsongdataset.com

Briot, Jean-Pierre , Gaetan Hadjeres, and Franc ois-David Pachet. "Deep Learning Techniques for Music Generation: A Survey." arXiv:1709.01620 (7 Aug 2019) https://arxiv.org/abs/1709.01620

Serra, Joan, Alvaro Corral, Marian Boguna, Martin Haro, and Josep Ll. Arcos. "Measuring the Evolution of Contemporary Western Popular Music." Nature Scientific Reports 521 (26 July 2012). https://www.nature.com/articles/srep00521

Poetry Machines: Empowering Creative Writers to Design DH Tools

Kyle Paul Booten
kyle.p.booten@dartmouth.edu
Dartmouth College, United States of America

Katy Ilonka Gero
katy@cs.columbia.edu
Columbia University, United States of America

Keywords: electronic literature, interface design, participatory design, natural language processing
Topics: English, North America, Contemporary, electronic literature production and analysis, Interface design, development, and analysis, Design studies, Literacy, composition, and creative writing

Digital Humanists often build tools based on their own understanding of what features users will need, whether intuited or derived from user studies. In the spirit of a key theme of DH2020---Public Digital Humanities---our presentation will consider what happens when we as researchers instead empower people to design their own DH tools according to their (perhaps idiosyncratic) purposes.

With our project, "Poetry Machines," we are recruiting individual poets who are not themselves programmers or machine learning practitioners to design their own artificially intelligent tools that will assist or challenge them in the middle of the writing process.

This research takes the approach of "participatory design" (Muller and Kuhn 1993). We have composed a design guide, a "boundary object" (Star and Griesemer 1989) providing a way for the poets to design their own poetry machines by combining various basic text processing functions and machine learning techniques into complex interfaces. We will then build these interfaces to their specifications, working with them collaboratively to further refine and develop designs (for instance, by presenting options of how a particular wished-for feature could be implemented).

A few possible functions that such machines could perform (the first two drawing from our own research [Booten 2019; Gero and Chilton 2019]): provide stylistic advice, e.g. alert the writer when she has used highly common lexico-semantic patterns or descriptors (e.g. 'glowing' to describe 'moon'); replace a word with a synonym that is characteristic of a particular author; whenever the writer pauses for too long, complete the line with a neural network

Our design guide provides the poet with substantial flexibility in terms of how the Poetry Machine is activated (for instance, randomly vs. after every line), how it intervenes (for instance, adding to the poem or displaying information in a sidebar), and how machine learning models are trained (for instance, a neural network trained on Romantic vs. modernist poetry).

Our research goals are to determine: 1) whether different sorts of poets (e.g. formalists vs. more experimental writers) make different design choices, 2) whether the machines meet their design expectations, and 3) whether this process provides the poets opportunity to develop critical perspectives on machine learning itself. Having acquired institutional approval to solicit poets for formal research, we are in the process of engaging poets in the participatory design process, and our presentation will offer preliminary findings.

This work intervenes in several DH fields: Though we are inspired by key examples of aesthetically-experimental hermeneutic interfaces (e.g. Clement 2008; McCurdy et al. 2016), our research will be of interest to DH tool designers more broadly, especially those interested in "bespoke" tools (Lin 2012); "Electronic literature" is part of DH (Rettberg 2015; see also Berkman 2017). Programmer-poets often design their own tools for algorithmic co-writing (e.g. Johnston's [2019] neural-network assisted ReRites). Our project aims to make this "e-lit" practice more widely available; This research connects to conversations about using DH to expand algorithmic literacy (D'Ignazio and Bhargava 2018).

References

Berkman, Natalie. "Digital Oulipo: Programming Potential Literature." DHQ: Digital Humanities Quarterly, vol. 11, no. 1, 2017. http://www.digitalhumanities.org/dhq/vol/11/3/000325/000325.html

Booten, Kyle. "Toward Digital Progymnasmata." Proceedings of the 10th International Conference on Computational Creativity, edited by Kazjon Grace, Michael Cook, Dan Ventura, and Mary Lou Maher, Association for Computational Creativity, 2019.

Clement, Tanya E. "'A Thing Not Beginning and Not Ending': Using Digital Tools to Distant-Read Gertrude Stein's The Making of Americans." Literary and Linguistic Computing, vol. 23, no. 3, 2008, pp. 361-381.

D'Ignazio, Catherine, and Rahul Bhargava. "Creative Data Literacy: A Constructionist Approach to Teaching Information Visualization." DHQ: Digital Humanities Quarterly, vol. 12, no. 4, 2018. http://digitalhumanities.org/dhq/vol/12/4/000403/000403.html

Gero, Katy Ilonka, and Lydia B. Chilton. "How a Stylistic, Machine-Generated Thesaurus Impacts a Writer's Process." Proceedings of the 2019 Conference on Creativity and Cognition, edited by Steven Dow, Mary Lou Maher, Andruid Kerne, and Celine Latulipe, Association for Computing Machinery, 2019.

Johnston, David Jhave. ReRites. Montreal, Anteism, 2019.

Lin, Yu-wei. "Transdisciplinarity and Digital Humanities: Lessons Learned from Developing Text-Mining Tools for Textual Analysis." Understanding Digital Humanities, edited by David M. Berry, Palgrave Macmillan, London, 2012, pp. 295-314.

McCurdy, Nina, et al. "Poemage: Visualizing the Sonic Topology of a Poem." IEEE Transactions on Visualization and Computer Graphics, vol. 22, no. 1, 2015, pp. 439-448.

Muller, Michael J., and Sarah Kuhn. "Participatory Design." Communications of the ACM, vol. 36, no. 6, 1993, pp. 24-28.

Rettberg, Scott. "Electronic Literature as Digital Humanities." A New Companion to Digital Humanities, edited by Susan Schreibman, Ray Siemens, and John Unsworth, Wiley, 2015, pp. 127-136.

Star, Susan Leigh, and James R. Griesemer. "Institutional Ecology, Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." Social Studies of Science, vol. 19, no. 3, 1989, pp. 387-420.

Championing Open Interactions with RERUM; Focusing and Preserving an Accessible Scholarly Conversation Across Disciplines

Patrick Miles Cuba
patrick.m.cuba@slu.edu
Saint Louis University, United States of America

Bryan J Haberberger
bryan.j.haberberger@slu.edu
Saint Louis University, United States of America

Keywords: annotation, IIIF, interoperability, design, API
Topics: Global, Europe, English, North America, 5th-14th Century, 18th Century, Contemporary, Interface design, development, and analysis, systems and information architecture and usability, Computer scienceHumanities computing

Rerum is a completely discrete set of services for creating and manipulating not only Web Annotations, but also IIIF documents and related objects. Our current services allow for generic interoperability and discoverability through the "unseen" services, coaxing the interface designer to intentionally shape the screen around the purpose of the user experience. In this way, the data model cannot be proprietary, removing the need for data import and export-interoperability has been replaced with simple operability.

Investigation of reliability in public history documents from 18th century

Cristina Vertan
cristina.vertan@uni-hamburg.de
University of Hamburg, Romania

Walther v. Hahn
University of Hamburg, Romania

Keywords: vagueness, uncertainty, ontology, digital history
Topics: Europe, English, 5th-14th Century, 15th-17th Century, 18th Century, data modeling, information retrieval and querying algorithms and methods, Humanities computing, Central/Eastern European Studies

In this paper we will report about both computer-based and hermeneutic approaches used for investigating multilingual documents and show how both can be combined for historical investigations. We will present the developed knowledge base a fuzzy OWL Ontology covering concepts, places and persons from the ottoman empire and neighbour countries, used as backbone for investigating reliability of public history documents.

Explore Plato's Cave in Virtual Reality

Julia DeWall Uhr
juuh6005@colorado.edu
University of Colorado, United States of America

Keywords: VR, Plato, Philosophy
Topics: Europe, English, BCE-4th Century, public humanities collaborations and methods, virtual and augmented reality creation, systems, and analysis, History, Philosophy

This lightning talk will introduce Escape Plato's Cave, an interactive virtual reality experience that transports the user into Plato's famous allegory about the power and peril of education and discuss it as an example of the potential uses of VR for both multimodal pedagogy and public philosophy.

"What Our Favorites Can Tell Us: A Computer-Directed Reading, Topic Modeling and Trend Analysis Investigation into 2,000 New York TimesBestsellers and Genre Fiction Winners from 2005 to 2016"

Morgan Lundy
lundym@email.sc.edu
University of South Carolina

Keywords: topic modelling, New York Times Bestsellers, trend analysis, text mining
Topics: English, North America, Contemporary, natural language processing, text mining and analysis, Library & information science, Literary studies

This study investigates what content and themes appear "at scale" in recently published literature, specifically New York Times bestsellers and genre fiction winners between 2005 and 2016. Topic modelling, a form of natural language processing and statistical analysis that gleans "topics" or related word groups by looking at the probability that words occur in the same documents within a corpus, is applied in this study to investigate a dataset of over 2,000 novels at the sentence level. This investigation embodies a DH sense of "play" by co-creating with computer (in this case, a statistical model) to identify both familiar and "latent" topics. The aim of this content analysis is to identify significant themes that resonate with recent social and political issues like the depiction of gun violence across genres, representation of minority characters, health-care related content and in tracking how the content of popular texts follows trends across a decade.

Typicality in the U.S. Novel

Mark Algee-Hewitt
Stanford University, United States of America

Erik Fredner
fredner@stanford.edu
Stanford University, United States of America

Keywords: typicality, U.S. literature
Topics: English, North America, 19th Century, 20th Century, text mining and analysis, Literary studies

Do literary critics know what a typical novel is like? Frederic Jameson all but claims to when he describes an "unexpected...modification" in a character's life as something that would "normally generate a properly Utopian narrative" (Jameson 184, emphasis added). Critics may reject the idea of any given novel's typicality, but paradoxically rely on a conception of what the typical novel does as a heuristic to make aesthetic judgments. We know when our expectations have been undercut, but we focus on what that undercutting does, rather than the origins of the expectation. This project shifts the critical focus from the former to the latter.

Where Víktor Shklovsky looks for a type of the novel —finding it in Tristram Shandy, the "most typical novel of world literature"—we seek typicality in the no less exaggerated form of the average (Shklovsky 52). In an essay on Josephine Miles's early distant reading, Brad Pasanek quotes an objection made by a reviewer to Miles's results that captures the tension between

the critical and the computational senses of typicality: "'If Quarles is "most typical," in any way at all, of this set of poets,' which includes Donne, John Milton, and John Dryden, 'I am immediately convinced that typicality is not a fruitful thing to investigate'" (Pasanek 369). We argue that reinvigorating the discourse of typicality through operationalization as Miles did forces a necessary confrontation between its meanings in quantitative and qualitative discourse.

We present a series of experimental approaches to the question of literary typicality. What, for instance, is the typical novel about? Using the Gale Corpus of American Fiction, which contains more than 18,000 texts published in the United States between 1774 and 1920, we analyze its 2,000 most frequent nouns as a proxy for the subjects of the novels. We then calculate the variance of each word in both concurrent and transhistorical groups to study how centrally embedded each individual text is within its literary field. We experiment with different measures of similarity-Euclidean distance, cosine similarity, and Kullback-Leibler divergence-to analyze which novels and which portions of novels in an individual author's corpus are least unlike the others in the set. This analysis challenges the critical shortcut, widespread at least since Auerbach's Mimesis, of identifying a given passage as "typical" of a given author. Finally, we use t-stochastic neighbor embedding to plot the relationships among all of the novels in the corpus, and analyze those at the center of the distribution, focusing on one novel that is more centrally located than any other. The same experiments are repeated using verbs: What does the typical U.S. novel do?

Through each of these experiments, we offer provisional operationalizations as provocations to think through the problems posed by literary typicality. If readers bring expectations to novels, those expectations come from a range of cultural forces, most salient of

which is the novel itself. Are those expectations justified by the predominant features of other novels? Or, to tweak Shklovsky's formulation, how typical is the average novel?

Bibliography

Jameson, Fredric. The Political Unconscious: Narrative as a Socially Symbolic Act. Routledge, 2013.

Pasanek, Brad. "Extreme Reading: Josephine Miles and the Scale of the Pre-Digital Digital Humanities." ELH, vol. 86, no. 2, June 2019, pp. 355-85. Project MUSE, doi:10.1353/elh.2019.0018.

Shklovsky, Víktor. The Novel: An Anthology of Criticism and Theory 1900-2000. Edited by Dorothy J. Hale, John Wiley & Sons, 2005.

Digital Humanities Start-Up Grant Retrospective: Towards a Sustainability Index for Grant Funded DH Projects

Cal Murgu
cmurgu@ncf.edu
New College of Florida, United States of America

Atalay Kutlay
New College of Florida, United States of America

Keywords: sustainability, digital humanities, index, qualitative, NEH
Topics: English, North America, Contemporary, digital research infrastructures development and analysis, sustainable procedures, systems, and methods, Library & information science

Sustainability is an important topic for digital humanities practitioners and funding agencies. As the recent NEH sponsored "Sustaining DH" series illustrates, strategies for sustainable development, dissemination, and preservation are critical if the "making" side of digital humanities are to persist. This paper details an ongoing attempt to develop a "sustainability index" for grant funded digital humanities projects. For the purposes of this paper, we define sustainability as the quality of being accessible/operable following a granting period. We focus on operability as our sustainability index is based on a qualitative analysis of projects that received NEH Digital Humanities Start-Up Grant funding from 2010-2017, specifically under the product type of "Web Resources" or "Computer Programs." Based on our analysis, we aim to show which criteria increase the likelihood of a sustainable project as well as the criteria that inhibit a project from remaining sustainable.

The Shakespeare and Company Project

Natalia Ermolaev

nataliae@princeton.edu
Princeton University, United States of America

Rebecca Koeser

rebecca.s.koeser@princeton.edu
Princeton University, United States of America

Keywords: software development, design, collaboration, data publishing, interface
Topics: Europe, English, 20th Century, data publishing projects, systems, and methods, software development, systems, analysis and methods, Humanities computing, Literary studies

The Shakespeare and Company Project is one of the longest-running projects at the Center for Digital Humanities at Princeton (CDH). Based on archival materials from the Sylvia Beach Papers at Princeton's Firestone Library, the project recreates the world of the Lost Generation by detailing what members of Beach's lending library read and where they lived. The full site launched May 2020, and data on library members, books, and events will be published soon.

The Shakespeare and Company Project is an exemplar of CDH's commitment to critical adoption of best practices (Koeser, 2019) and innovation in custom software. Our process leverages rigorous testing, including unit testing (over 95% coverage for Python code), automated accessibility testing with pa11y, and manual feature and usability testing. As a project with both popular and scholarly appeal, we've designed a site that foregrounds the complex detail and idiosyncrasy of Beach's record-keeping practices, but is welcoming and accessible, and also fully responsive on mobile devices. Through collaborative and iterative data modeling, we developed a relational database that allowed the research team to aggregate data , which helped to identify and disambiguate members, and to associate events from different archival sources. This project also includes a sophisticated solution for storing and querying partially-known dates in a SQL database (Koeser, 2019).

Our lightning talk will demonstrate aspects of the public interface as well as the administrative backend, provide a brief overview of the published data, and discuss some of the scholarly and technical challenges we encountered.

References

Shakespeare and Company Project. Center for Digital Humanities, Princeton University, 2020. https://shakespeareandco.princeton.edu/.

"Data Export." Shakespeare and Company Project, version 1.1.0. Center for Digital Humanities, Princeton University. April 13, 2020. http://shakespeareandco.princeton.edu/about/data/. Accessed June 15, 2020.

Koeser, Rebecca Sutton, Nick Budak, Gissoo Doroudian, Benjamin W. Hicks, and Xinyi Li. 2020. Princeton-CDH/mep-django: v1.0. Zenodo. https://doi.org/10.5281/zenodo.3834179.

Koeser, Rebecca Sutton. 2019. "Best Practices?" Panel "The State of Digital Humanities Software Development." presented at the ACH2019, Pittsburgh, July 26. https://rlskoeser.github.io/2019/07/26/best-practices/.

Koeser, Rebecca Sutton. 2019. "Coding with Unknowns." Center for Digital Humanities at Princeton University. December 5, 2019. https://cdh.princeton.edu/updates/2019/12/05/coding-unknowns/.

Computational access to library's digital collections

Yoo Young Lee

ylee4@uottawa.ca
University of Ottawa Library, Canada

Keywords: Computational access, web archiving, WARC
Topics: English, North America, Contemporary, digital archiving, digital libraries creation, management, and analysis, First nations and indigenous studies, Library & information science

Born-digital or digitized resources enable researchers to apply computational methods to various research topics in the field of digital humanities (Klingenstein, Hitchcock, & DeDeo, 2014; Nanni, Dietz, & Ponzetto, 2017). There have been tutorials and workshops on how to apply them with multiple tools like R or Python targeted for DH researchers (Unsworth, 2009; Mäkelä, 2019; Mullen, 2018; The Programming Historian, n.d.; The Digital Humanities Summer Institute: Technologies East 2020, n.d.). In addition, there is an initiative, Always Already Computational: Collections as Data to develop a strategic direction and guide libraries and cultural heritage institutions to provide collections as data for researchers so that they can leverage computational methods (Padilla, Allen, Frost, Potvin, Russey Roke, & Varner, 2019).

In this lightning talk, I will share my journey to making the NAHO (National Aboriginal Health Organization)

WARC (web archive format) file computationally accessible. The University of Ottawa Library initiated a web archiving project to preserve entire NAHO web content which is currently accessible only via Wayback Machine. There is no easy way to extract data from the NAHO collection which prevents researchers to apply computational methods in research tools like R, Python, or Archives Unleashed toolkit. I will also talk about processes, challenges, and resources in order to provide computational access to library collections with an example of the NAHO WARC file.

This lightning talk is targeted for DH or librarians, but open for all and they can learn the importance and processes of how to make digital collections computationally accessible.

References

Unsworth, J. (2009). Computational methods in humanities research. Retrieved from http://www.people.virginia.edu/~jmu2m/florence.09.html

Klingenstein, S., Hitchcock, T., & DeDeo, S. (2014). The civilizing process in London's Old Bailey. Proceedings of the National Academy of Sciences, 111(26), 9419-9424. doi: 10.1073/pnas.1405984111

Maemura, E., Becker, C., & Milligan, I. (2016, December). Understanding computational web archives research methods using research objects. In 2016 IEEE International Conference on Big Data (Big Data) (pp. 3250-3259). IEEE. doi: 10.1109/BigData.2016.7840982

Mäkelä, E. (2019). Introduction to methods for digital humanities. Retrieved from https://jiemakel.gitbook.io/meth4dh/

Mullen, L. A. (2018). Computational historical thinking: With applications in R. Retrieved from https://dh-r.lincolnmullen.com/

Nanni, F., Dietz, L., & Ponzetto, S. P. (2017). Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts. Digital Scholarship in the Humanities, 33(3), 612-620. doi: 10.1093/llc/fqx062

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019). Final Report --- Always Already Computational: Collections as Data. https://doi.org/10.5281/zenodo.3152935 (project site: https://collectionsasdata.github.io/)

The Digital Humanities Summer Institute: Technologies East 2020. (n.d.) Retrieved from https://dhsite.org/

The Programming Historian. (n.d.). Retrieved from https://programminghistorian.org/en/

Weingart, S., & Jorgensen, J. (2012). Computational analysis of the body in European fairy tales. Literary and Linguistic Computing, 28(3), 404-416. doi: 10.1093/llc/fqs015

MONA from public art to our art with a mobile app

Lena Krause
lena.krause@umontreal.ca
Université de Montréal, Canada

Keywords: mobile app, public, art, cartography, gamification
Topics: English, North America, Contemporary, mobile applications development and analysis, public humanities collaborations and methods, Art history, Cultural studies

The term "public art" commonly designates the sculptures, monuments, and murals that adorn our cities. Although this art form can be ignored or go unnoticed (Veyne, 1988), recent events have led to many interactions with it. On the one hand, the Covid-19 pandemic has brought attention to the subject of access to culture. Due to confinement and imposed social distancing measures, citizens rediscovered their urban settings and thus renewed their interest for public art. On the other hand, the Black Lives Matter movement has revived debates over which artworks should be publicly displayed. The statue of the slave trader Edward Colston in Bristol is a historical symbol of white dominant colonial history. Far from an act of gratuitous "vandalism," its toppling by protesters can rather be understood as an outcry of citizens whose voices have not been listened to. They therefore took matters into their own hands and, as a result, took part in the curation of their city's public art collection. The destruction of art is recurrent in history, and its motives range from political upheavals to a variety of misunderstandings (Gamboni, 2007). Yet, even when confronted with such "monumental controversies" (Senie, 2019), decisions and discourses about public art are mostly generated by local government officers, politicians, cultural workers, or journalists. Although citizens are sometimes invited to be part of public art commissions, they usually lack the opportunity to voice their opinion. They are considered to be the recipients of public art policies whose aims are to elevate their culture and spirit.

This top-down approach attributes a passive role to spectators, and such conception permeates onto the tools that promote public art. Our research into contemporary mobile applications dedicated to this art form reveals that most apps provide limited interactivity. They are digital equivalents of traditional guidebooks, enhanced with geolocation and audio-guides (Chicago

Public Art, 2015, Otocast, 2019). More responsive, the ARTsansRDV app (Geneva, Switzerland, 2015) adapts the audio content to the user's location. Nevertheless, these digital tools provide selected information. They instruct users on how to perceive the artworks from the perspective of cultural elites, leaving little room for individual perception and interpretation. This one-sided approach explains the lack of participatory features. The AtlasMuseum's app (now deprecated) invited to greater involvement as the project aims to make an inventory of public artworks through crowdsourcing (Pringuet, 2017). However, this active stance requires previous knowledge and focuses on informative documentation.

The underlying goal of these institutions has been to remove the geographical and financial divide, and thus democratize access to art (Donnat and Tolila, 2003: 8). However, even after "extracting culture from its symbolic and architectural shell" (Ghebaur, 2013: 709), interacting with it remains difficult. Whether it is in the choice of what is represented or in the way one should understand it, citizens are subjected to a sort of "institutional tyranny." The MONA app was created to rethink public art, as a reminder to define it not only through location or funding, but also as a common property. Its name derives from a pun, with mon (my, the possessive pronoun) placed in front of the letter "a" for art and app alike. The app has two main goals. It confers citizens the right to have their own interpretation of public artworks. It also symbolically grants each individual their share of responsibility and accountability in their artistic environment.

This free and open source app aligns basic smartphone functionalities with art discovery through gamification, personalization, and imagination. Each artwork is mapped and users are invited to search for them as in a life-sized treasure hunt. Once discovered, they interact with the art via mobile photography. They capture their own perspective of the artwork, voice their opinion through a short comment and produce a rating of their experience. Each discovery adds to their personal collection, which becomes over time a catalog illustrated by their own photographs. As the collection expands, the user receives badges as a reward and encouragement to pursue their exploration. No previous artistic knowledge is required as the app accompanies users in their discoveries, observations and interpretations of the artworks. Conversely, group activities (temporarily online) are organized to help those less confortable with the use of technologies to apprehend the workings of the app.

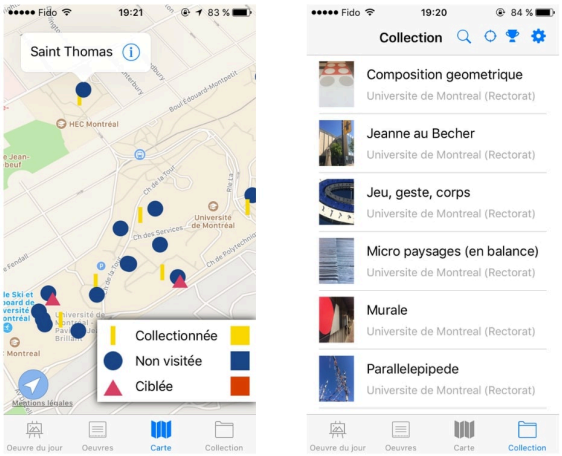


Fig.1

Fostering the emergence of the public's opinions has a civil and an academic purpose. The photographs, comments, and ratings provided by the users are collected anonymously on university servers (under our privacy policy). These will allow us to study, for the first time, how people see, feel and think about public art. Research and publications derived from the app aim to share these responses, giving authority to the participation of citizens. Now more than ever, people express an active interest in the historical and cultural representations that shape their city's image and identities. MONA users are therefore encouraged to participate in sharing their discoveries and opinions on social media. All ? including artists, historians and political figures, but also families, local dog-walkers and curious teenagers ? get to share their thoughts on historical figures commemorated in their neighbourhood parks. This is a particularly vivid example for the MONA team since the nearest artwork from our headquarters is the bust of Christopher Columbus (Armand De Palma, 1976, Montreal's public art collection). Altogether, such bottom-up structure creates opportunities to hear from a diversity of people, which is not only valuable for our society but also provides a unique and innovative research field.

Our singular take on public art is the result of a practical, "real-life" approach that corresponds to DH methods for a knowledge base grounded in making and experimentation (Burdick et al. 2012: 57). The app's metadata is provided by Montréal and Québec's open data platforms. The iOS and Android apps are each developed in the native languages, Swift and Java respectively, in order to meet industry standards and user expectations. First prototyped in the class room, MONA has become a collaborative project uniting students in art history, computer science, design, and

more. Combining a research project (Art+site) with the work of a non-profit organisation (La Maison MONA), we offer new approaches to the study of public art while also encouraging rich and critical exchanges between all citizens. As digital alternatives flourish in the cultural field, we endeavour to offer physical access to culture and to use the digital realm to create a space of expression for new and unheard voices. Each contribution makes it a little more our art, aiming to bridge social and cultural divides by taking into account our society's diversity and complexity.

Enhancing Community through Open DH Website Design

Rennie Mapp
mapp@virginia.edu
U of Virginia, United States of America

Christian Howard-Sukhil
cfh008@bucknell.edu
Bucknell University

Keywords: design, website development
Topics: English, North America, Contemporary, Interface design, development, and analysis, project design, organization, management, Design studies, Informatics

Our lightning talk offers solutions to some shortcomings in communication about DH projects and undertakings on university campuses, particularly through the development of institutional DH websites. By an "institutional DH website," we mean a community website, hosted by a given university or institution, that is explicitly devoted to the advancement, support, and promotion of DH work collectively.[1] A number of previous attempts to establish institutional DH websites have failed,[2] and there is a growing need to understand how we can sustainably create and maintain such sites in a way that meets the diverse needs of DH scholars. To this end, we offer an alternative approach for creating such communal sites that is designed for specific communities. More than merely providing a definition of DH and a set of resources for those interested in the field, institutional DH websites can beneficially act as community hubs for DH practitioners by showcasing live projects and encouraging interdisciplinary collaboration. While there is no one-size-fits-all solution, an open development process can help scholars and DH staff who face long-standing DH challenges around methodological innovation, data reproducibility, reinvention of the wheel, and the balance of technical and humanistic priorities. In particular, we offer a user-focused development process for DH websites, which emphasizes the identification and enhancement of human networks and communities of practice. At the most basic level, user-focused design starts with a needs assessment of the website's primary audience and is refined through attention to typical user needs and exemplary uses throughout the project's lifecycle in order to maintain an active user community. This stands in contrast to the design of institutional DH sites as a means of cataloguing the services or offerings at a specific institution. At the structural level, user-focused design for institutional DH sites foregrounds open access and accessibility by thinking about these concerns throughout the design process (rather than as a last-minute add-on).[3]

The two speakers have designed institutional DH websites at the University of Virginia and Bucknell University,[4] each employing a different platform (Drupal and WordPress); nonetheless, both sites promote similar design philosophies. The talk will model how institutions can create similar sites designed for their own communities with an eye toward developing appropriate use cases and sustainability practices.

[1] For instance, see the website for Duke University's Digital Humanities Initiative (https://digitalhumanities.duke.edu/) or Stanford's DH website (https://digitalhumanities.stanford.edu/).

[2] See, for instance, the DiRT Directory (https://blogs.shu.edu/digitalhumanities/2018/04/25/dirt-directory/), Project Bamboo (https://www.projectbamboo.org/), and MIT's Hyperstudio (http://hyperstudio.mit.edu/).

[3] For specific examples of what it means to design websites with open access and accessibility in mind, see principles developed by the Fair Open Access Alliance (https://www.fairopenaccess.org/the-fair-open-access-principles/) and the Web Accessibility Initiative (https://www.w3.org/WAI/fundamentals/accessibility-principles/).

[4] See DH@UVA (https://dh.virginia.edu/) and DH@Bucknell (http://dhbucknell.scholar.bucknell.edu/).

Structures of Thought: Common Readers and Data Modeling Annotations

Rebecca Munson
rmunson@princeton.edu
Princeton University, United States of America

Keywords: annotation, data modeling, book history, text encoding, database Topics: Europe, English, North America, Australia/
Oceania, 15th-17th Century, Contemporary, annotation structures, systems, and methods, data modeling, Book and print history, Literary studies

Common Readers is a digital initiative dedicated to gathering and analyzing annotations in early modern playbooks through a custom-designed relational database. The corpus of the pilot phase consists of drama in English, 1550-1660. The goal is to foster a community of researchers to contribute to a growing dataset of annotations that will, ultimately, be part of a public-facing site. To date, there have been two major trends in annotation scholarship: the study of remarkable men (e.g. John Dee) and the study of remarkable books (e.g. a copy of Shakespeare's First Folio held at Meisei University). DH work on annotations has followed suit, striving to make annotations by particular readers in particular books easily searchable.Common Readers stands in contrast to both these scholarly trends in its potential to illuminate the reading habits of the less extraordinary, anonymous men and women who read, marked, and commented upon texts. This talk will introduce the audience to Common Readers and solicit feedback on the existing annotations data module to improve its implementation, accessibility, and appeal to a variety of users. It considers the construction of the data schema through particular use cases as an opportunity to reflect on the process of structuring data, and in particular on the practical and scholarly implications of the choice between an hierarchical encoding structure like TEI and a relational database. (For an example of the former, see the Archaeology of Reading project; for an example of the latter see Derrida's Margins.)Moreover, this talk will consider the way in which annotated books--as multivalent, hybrid objects--challenge many traditional structuring principles in and outside of DH. As an inquiry into the best approach to annotation studies, Common Readers raises broader questions about the ontology of annotated documents that engage many disciplines and subfields, offering potential applications for researchers who are curating data about complex material objects, designing platforms for capturing annotations, or practicing scholarship related to object-oriented ontology. In this emphasis on data structures, Common Readers seeks community guidance on its approach to making highly particular data not only publicly shareable but widely useful. Originally, the plan was to provide the (future) public site with an export function, allowing users to work with project data in whatever way best suited their needs, and to deposit versioned datasets for general use in a repository like Zenodo. Both practices are likely to benefit other early modernists, but probably not a general audience. While the appeal of Linked Open Data (LOD) is strong) the best way in which to structure such highly specific data (derived from a very particular use case) in a way that would comply with recognized standards and be widely machine-readable remains unclear. The project team welcomes advice in establishing these best practices, as well as for redesigning the annotations data module itself.Addendum: The disruption caused by COVID-19 has delayed finalizing the annotations data module discussed above. If you would be interested in reviewing it in the future, please email Rebecca Munson (rmunson@princeton.edu).

Argumentation Mining in Spoken Georgian Political Language

Mariam Matiashvili
mariammatiashvili26@gmail.com
Iv. Javakhishvili Tbilisi State University, Georgia

Keywords: Argumentation, Annotation, Text mining, Spoken languages, Georgian
Topics: Europe, English, Contemporary, annotation structures, systems, and methods, text mining and analysis, Linguistics

The research concerns argumentation mining, specifically issues related to its theoretical framework and computational modeling.

The research goals are: (a) A linguistic analysis of argumentative structures in spoken Georgian political

language. (b) The study of argumentation mining in spoken language and the elaboration of a Database Processing/Preparation System using those results.

The research data comes from the Georgian Political Spoken Language Corpora (http://gnc.gov.ge/gnc/page?page-id=gnc-main-page). Currently, the research is at the stage of lexical analysis of the text, that is, attempting to answer the question of what kind of lexicon is used in Georgian Spoken Argumentation Language.

Finally, I aim to create an Annotation Scheme that can be a base for (a) the Arguing Corpora (b) a Labeled Database for Machine Learning training that aims at Argumentation Mining.

Short Presentations

Booth Tarkington, Blindness, Dictation, and the Durability of Style

David Lowell Hoover
david.hoover@nyu.edu
New York University - Main Campus, United States of America

Keywords: style, dictation, literary composition
Topics: English, North America, 19th Century, 20th Century, attribution studies and stylometric analysis, Literary studies

When Booth Tarkington suffered from severe vision problems and temporary complete blindness in 1929, he began to dictate almost all of his literary works, a practice he continued even after regaining good eyesight in 1931. This submission investigates the possible effects of this change in mode of composition on Tarkington's style. The tentative conclusion is that there is almost no evidence of any effect, which suggests that authorial style can be quite durable in the face of a complete change in the way the author produces his text.

Testing Rolling.Classify

David Lowell Hoover
david.hoover@nyu.edu
New York University - Main Campus, United States of America

Keywords: Rolling Classify, collaboration, simulation
Topics: Europe, English, North America, 19th Century, 20th Century, Contemporary, attribution studies and stylometric analysis, Humanities computing, Literary studies,

Rolling.Classify is a recently developed tool for studying collaboration (Eder, Rybicki, and Kestemont 2016; Eder 2016) that builds on earlier work that tested successive overlapping sections of texts (van Dalen-Oskam and van Zundert 2007, Burrows 2010, Hoover 2012).

The power and ease of use of Rolling.Classify (and its related Rolling.Delta) have led to several studies based on various kinds of texts.. Rigorous testing of this new method on problems with known solutions seems especially important because its results vary greatly with the choice of classification method other parameters. I will begin with simulated collaborations comprising text sections of varied lengths assembled to model different kinds of collaboration. I will then test collaborations with known contributions by the authors, and finally some in which no clear evidence of the nature of the collaboration exists.

Ethics and responsibilities of open access - lessons learned from the Wikipedia project of the Atikamekw First Nation

Nastasia Herold
nastasia.herold@uni-leipzig.de
University of Leipzig, Germany Wikipetcia Atikamekw Nehiromowin, Canada

Thérèse Ottawa
thereseottawa@gmail.com
Wikipetcia Atikamekw Nehiromowin, Canada

Keywords: open access, wikipedia, atikamekw, first nations, collaborative
Topics: English, North America, Contemporary, digital access, privacy, and ethics analysis, open access methods, First nations and indigenous studies, Linguistics

Abstract by Nastasia Herold & Thérèse Ottawa

Ethics and responsibilities of open access - lessons learned from the Wikipedia project of the Atikamekw First Nation

With 97,9% (cf. INAC 2019), in Canada, the Atikamekw First Nation has the highest percentage of people who speak their native language (Atikamekw) at home. The Atikamekw live in Quebec, Canada, and have a population of 8,000 people in three communities.

In 2013, Nastasia Herold, one of the authors of this paper did a field study in Manawan, one of the three Atikamekw communities, for a research on the local bilingualism (Atikamekw and French). Despite the vitality of the Atikamekw language, a survey and

interviews showed that francization and language change are processes noticed by all living generations of the Atikamekw (cf. Herold 2020: N.N.).

Communication takes place more and more often digitally (cf. Reichert 2017: 26-27), and this in written language rather than orally. Atikamekw has a standardized orthography since 1994 (cf. Dinnison 21997) and is taught in Manawan's primary school as a first language and as medium of alphabetization. However, Herold's (2019: 103) research in 2013 showed that the Internet contained no written text in the Atikamekw language, the Atikamekw used the Internet mainly in French. This is why a school project at Manawan's secondary school was initiated in 2013 in order to create a Wikipedia site in the Atikamekw language.

Many lessons have been learned during the collaboration of academics, teachers, pupils, local language experts and other local voluntary contributors. The lessons we would like to focus on in this presentation are the lessons learned when implementing cultural knowledge to an open access platform. Open access is the free provision of (scientific) texts on the web without restriction of use (cf. Kohle 2017: 203), and this free provision has advantages and disadvantages.

We will give four examples which show that the Atikamekw made sure that the knowledge was published respecting their own principles, beliefs and tradition. These four examples and the development of the project show how it is important that the community itself dictates rules that have to be respected when publishing their knowledge under open access.

Finally we will answer Rehbein's and Thies' (2017: 355) question they developed as a schema for questions of responsibilities and ethics of a specific project: Who (1) is responsible for what (2) to whom (3) before which instance (4) according to which standards (5)?

References

Dinnison, Bonnie (21997): Guide orthographique de la langue atikamekw. La Tuque: Atikamekw Sipi.

Herold, Nastasia (2019): "Kulturbedingte Herausforderungen für schulische und gesellschaftliche Teilhabe in indigenen Reservaten Kanadas: Das Beispiel des Atikamekw-Dorfes Manawan in Québec", in: Jahr, David / Kruschel, Robert (eds.): Inklusion in Kanada. Internationale Perspektiven auf heterogenitätssensible Bildung. Weinheim: Beltz Juventa 92-106.

Herold, Nastasia (2020): "La situation linguistique bilingue dans la communauté atikamekw de Manawan - La consolidation d'une langue minoritaire", in: Recherches amérindiennes au Québec, N.N. [submitted].

INAC (26/09/2019): Tribal Council Detail. Atikamekw Sipi - Conseil de la Nation Atikamekw (01/10/2019).

Kohle, Hubertus (2017): "Digitales Publizieren", in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler 199-205.

Rehbein, Malte / Thies, Christian (2017): "Ethik", in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler 353-357.

Reichert, Ramón (2017): "Theorien digitaler Medien", in: Jannidis, Fotis / Kohle, Hubertus / Rehbein, Malte (eds.): Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler 19-34.

CR/10 Website as a Digital Public Humanities (DPH) Site

Rongqian Ma
rom77@pitt.edu
University of Pittsburgh, United States of America

Keywords: Public digital humanities, user experience design, CR/10, cultural revolution
Topics: Asia, English, North America, 20th Century, Contemporary, public humanities collaborations and methods, user experience design and analysis, Design studies, Library & information science

Initiated by Mao Zedong (1893-1976) and lasting from 1966 to 1976, the Cultural Revolution in China has been defined as a 10-year disaster by the Chinese Communist Party (CCP), which caused civil unrest to the party, the state, and the people (CCP, 1981). Forty years after the end of the Cultural Revolution, with the intention to promote the public remembrance and discussion of this significant period of Chinese history, the East Asian Library at the University of Pittsburgh launched CR/10 (University of Pittsburgh, 2019), a digital oral history project that aims to collect and preserve authentic memories of the Cultural Revolution through 10-minute semi-structured interviews. Using the technique of snowball sampling for interviewee selection, the project has thus far collected more than 300 interviews with ordinary

people from different generations, geographies, social and educational backgrounds who experienced the incident or learned about it from family, school, or other resources. Among all the interviewees, some currently live in the U.S. while some in China, whose everyday experiences post the Cultural Revolution may have exerted impacts on their individual memories. Most interviews were conducted in mandarin Chinese and were then translated into English as subtitles by the CR/10 team. All the interviews were video-recorded and made open-access on the interactive CR/10 website. The website consists of four major components: an introduction to the project, a trailer for the project that showcases interviews, a timeline and a map of China demonstrating the temporal and geographical distributions of the videos in the collection.

In this paper, I treat the CR/10 website as the object of study and demonstrate its value as a digital public humanities site bridging the gap between the academy and the general public. Cox and Tilton (2019) defined the term "digital public humanities (DPH)" as practices that "facilitate reflection and collaboration with participants outside of the academy through digital theories and technologies" (p. 130). Focusing on interactive and mindful design (Drucker, 2013), the CR/10 website invites the public to participate in, as well as contributing to, developing a diversified and multifaceted understanding of the Cultural Revolution. In this study, I present a series of user experience research conducted regarding the design of the website with 15 users outside of the academy, to examine the current usability as well as to identify further design possibilities of the website. Through techniques of semi-structured contextual interviews and focus group study, this research study aims to reflect on the reception and use of the CR/10 website as an interactive teaching and learning platform, rather than solely a repository of collective memories or database of historical information. More specifically, I examine how the timeline and map features of the website enrich interpretations of the Cultural Revolution, especially in terms of inspiring users to reflect upon the following questions: How could the Cultural Revolution be defined? What factors (e.g., geography, generation, family, class backgrounds, and education) influence impressions and memories of the Cultural Revolution? Extending from the user experience research, this study also proposes recommendations to improve the design of the CR/10 website to create a collective "memory atlas" (Cornell University Library, 2013; Forster, 1976) out of the video collection. Findings of this study contribute to facilitating academy-public collaborations in building DPH sites from the user perspective and a design-mediated approach.

References

Chinese Communist Party (CCP). (1981). Resolution on certain questions in the history of our Party since the founding of the People's Republic of China. Retrieved from http://en.people.cn/dengxp/vol2/text/b1420.html

Cornell University Library. (2013). Mnemosyne: Meanderings through Aby Warburg's Atlas. Retrieved from https://warburg.library.cornell.edu/

Cox, J., & Tilton, L. (2019). The digital public humanities: Giving new arguments and new ways to argue. Review of Communication 19 (2), 127-146.

Drucker, J. (2013). Reading interface. PMLA 128 (1), 213-220.

Forster, K. (1976). Aby Warburg's History of Art: Collective Memory and the Social Mediation of Images. Daedalus 105 (1), 169-176.

University of Pittsburgh. (2019). CR/10. Retrieved from http://www.culturalrevolution.pitt.edu/.

Features of Timelessness: Intermediate Report on a Quest for Stylistic Features that Mark Literary Canonicity

Joris J. Van Zundert
joris.van.zundert@huygens.knaw.nl
Department of Literary Studies, Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences. Amsterdam, The Netherlands

Raymond A. Mar
Department of Psychology, York University. Toronto, Canada

Karina van Dalen–Oskam
Department of Literary Studies, Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences. Amsterdam, The Netherlands Faculty of Humanities, University of Amsterdam. Amsterdam, The Netherlands

Emily Temple
Literary Hub

Isabel Bowman
Department of Psychology, University of Toronto. Toronto, Canada

Farzaneh Heidari
Department of Electrical Engineering and Computer Sciences, York University. Toronto, Canada

Ahn T.P. Nguyen
Department of Psychology, York University. Toronto, Canada

Keywords: stylometry, literature, features, canonicity
Topics: Europe, English, North America, 20th Century, Contemporary, attribution studies and stylometric analysis, Humanities computing, Literary studies

We report on our ongoing quest to establish a validated complex of stylistic features that act as markers for literary canonicity, in specific contexts. Currentely we present a stylometric analysis of literature investigating the stylistic markers that differentiate former bestsellers from fiction that remains popular across several decades using a Tfldf vectorization of texts and UMAP dimenision reduction approach. We find that especially a greater variation in sentence length is associated with the chances of a novel to remain popular.

Gmail's Smart Compose: A Critical Composit(ion)

Crystal Nicole Chokshi
crystal.chokshi1@ucalgary.ca
University of Calgary, Canada

Keywords: algorithmically-mediated writing, Smart Compose, critical algorithm studies, critical media studies, linguistic capitalism
Topics: English, North America, Contemporary, information retrieval and querying algorithms and methods, text mining and analysis, Communication studies, Media studies

While Google once merely monitored users' words, today the company literally writes them. This is thanks to Smart Compose, a word-prediction algorithm that Google has launched in Gmail and Google Docs. The algorithm depends on Google's meticulous recording and machine-reading of the personal data of an untold number of its 1.5 billion Gmail users, leveraging users' words and writing for the development of the technology. In this context, language is particularly vulnerable to corporate intervention and manipulation. As such, this presentation carefully considers Fréderic Kaplan's (2014) call to action: "through... the advent of algorithms as a new media, something is likely happen [sic] to language, and, although we are not yet sure what it will be, new tools must be built in order to understand this global linguistic evolution" (p. 62).

Responding to Kaplan's call, I report on experiments with Smart Compose in which I am manually transcribing more than 50,000 words from published texts and, subsequently, annotating and visualizing input to and output from the algorithm. These experiments are part of my larger doctoral project that seeks to locate the shifting "semantic coordinates" (Striphas, 2015, p. 398) of "language," "words," and "writing" in an algorithmic culture. Applying the framework of data colonialism (Couldry & Mejias, 2019), I argue that word-prediction algorithms such as Smart Compose must necessarily shift our understanding of these terms when words become data (Thornton, 2019) and writing becomes a datafied practice.

More broadly, I suggest that in place of the question posed by Siva Vaidhyanathan (2011) some years ago-"what do we gain and what do we lose by inviting Google to be the lens through which we see the world?" (p. 9)-we must, urgently and necessarily, ask this: what do we gain and what do we lose by allowing Google to offer the words through which we write the world?

References

Couldry, N. & Mejias, U.A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. Television & New Media, 20(4), 336-349. https://doi.org/10.1177/1527476418796632

Kaplan, F. (2014). Linguistic capitalism and algorithmic mediation. Representations, 127(1), 57-63. https://doi.org/10.1525/rep.2014.127.1.57

Striphas, T. (2015). Algorithmic culture. European Journal of Cultural Studies, 18(4-5), 395-412. https://doi.org/10.1177/1367549415577392

Thornton, P. (2019). Language in the age of algorithmic reproduction: A critique of linguistic capitalism [Unpublished doctoral dissertation]. Royal Holloway, University of London.

Vaidhyanathan, S. (2011). The googlization of everything (and why we should worry). University of California Press.

Transing DH: Adopting a Trans-Centric Approach to Building the Transgender Media Portal

Laura Horak
laura.horak@carleton.ca
Carleton University, Canada

Keywords: transgender, cinema, database, ethics, feminist
Topics: English, North America, 20th Century, Contemporary, database creation, management, and analysis, public humanities collaborations and methods, Film and cinema arts studies, Transgender and non-binary studies

Guided by the fields of intersectional feminist digital humanities and transgender studies, this talk explores our team's efforts to adopt a trans-centric approach to building a collaborative online tool to investigate and publicize films made by trans, Two Spirit, nonbinary, intersex, and gender-nonconforming people.

Theoretical intersections: cognitive poetics, cultural evolution, and distant reading in literary studies

Fabio Ciotti
fabio.ciotti@uniroma2.it
Università di Roma Tor Vergata, Italy

Keywords: distant reading, cognitive poetics, computational literary studies, cultural evolution, cultural analytics
Topics: Europe, English, 20th Century, Contemporary, text encoding and markup language creation, deployment, and analysis, text mining and analysis, Cognitive sciences and psychology, Literary studies

Since Franco Moretti coined the widely successful term "distant reading" (Moretti, 2000) quantitative /computational text analysis methods have gained a wide circulation in literary studies.

We can even speak of a distant reading school, nowadays. The diffusion of distant reading approaches has raised a lively debate (mostly in the North American context), and has attracted various criticisms, both from "traditional literary scholars" and self-critical adopters that can be subsumed into a threefold typology:

- Theoretical/Ideological: the intentional and qualitat ive nature of the literary domain is in principle irreducible to quantitative and computational methods; literature is not data (Marche, 2012) and literary criticism is not data analytics (Fish, 2012);
- Methodological: the (statistical/computational) models and methods adopted for literary analysis are wrong, inaccurate, and ultimately inadequate (Da, 2019);
- Pragmatical: the limits in the representativeness of the textual data set used in the analysis and the problems in defining the adequacy of its selection criteria (Bode, 2018).

Each of these kinds of criticisms would need a deep discussion and are strictly interconnected. For instance, in the well-known Da's articles, allegedly oriented presented as a replication failure study, the following excerpts makes apparent that the author has an a priori skepticism about the epistemological possibility of what she calls "computational literary studies:

There is a fundamental mismatch between the statistical tools that are used and the objects to which they are applied. (Da, 2019: 601)

[...] It may be the case that computational textual analysis has a threshold of optimal utility, and literature-in particular, reading literature well-is that cut-off point. (2019: 639)

I think that one of the main reasons underlying these more or less critical positions toward distant reading is the fact that it lacks sound and coherent rationales from the point of view of the theory: in fact, we can say that distant reading is the first methodology in literary studies that does not come with a theory of literature embedded in it, as it was for all of its predecessors. Consequently, all distant reading studies derive their theoretical frameworks and terms from theories in the literary domain that generally relies on the fundamental idea that literary texts can be explained only by the way of interpretation or if we prefer of hermeneutics.

The problem is that any literary interpretation based on quantitative, immanent, and purely formalist approach is subject to the theoretical criticism that was expressed by Stanley Fish in his harsh and seemingly ultimate criticism to stylistics in "What Is Stylistics, and Why Are They Saying Such Terrible Things About It?" (Fish, 1980). The point for Fish was not to criticize the methods per se, but the

possibility to extract meaningful literary interpretations directly from the simple linguistic facts, the idea of an "algorithmic interpretation" (Fish's words!), since interpretation always starts from a contextual and situated point of view that pre-defines the very objects of its actuation.

Many important scholars active in the field do believe that there is the possibility to reconcile traditional theories (of literature) with computational/ quantitative methods. Just to make a couple of examples of these consilience theses, we can cite Andrew Piper (Piper, 2018) and Michael Gavin (Gavin, 2018). Piper proposes a sort of computational hermeneutics, that integrates distant and close, and quantitative and qualitative readings. Michael Gavin argues that "vector semantics share a set of assumptions with literary critic William Empson, who devoted his career to explaining how poets played with words' many meanings". What is suspicious in these (and similar but often less intriguing and thought-provoking) calls to the reconciliation of the two poles is that their outcomes are either literary critically unsatisfactory or are self-contradictory, in that the hermeneutical and critical part of the discourse is self-standing, the critical arguments are logically independent from the results of the computational analysis.

On the base of these considerations, I think that distant reading simply cannot be considered a methodological innovation to be applied to our pre-existing theories of literary texts (in all their rhizomatic variants): it is necessary to find a suitable theory or framework where these methods can yield to interesting results. To make a substantial step in this direction we should first of all take seriously the notion of distant reading and abandon the idea of literature as either made of singular special individuals (the great or the small texts, amenable to interpretations by literary critics, or the big or small writers) or reduced to abstract ideal type, under the scrutiny of literary theoreticians with no clue with empirical evidence. This move would import also taking seriously the move from (or renounce to) interpretation to (embrace) explanation as the real aim of the scholarly inquiry.

On what theoretical basis, then, can we construct a notion of literature amenable to distant reading methods?

One possible direction to be explored as some scholars like Ted Underwood suggests (English and Underwood, 2016; Underwood, 2017), is that distant reading should fall inside the tradition of sociology of literature or history of ideas a la Nouvelle Histoire. Although there are many reasons to lean toward a sociological vision of literature as an optimal base for distant reading, I think that an even better theoretical framework is that of the cognitive and bio-evolutionistic approaches to literature and the cultural evolution studies. Cognitive poetics/narratology, and bio/evolutionary literary studies have been two of the most interesting waves of innovation in the literary field of the last 30 years and are now established field of inquiry. With different graduation depending on the authors, they have advocated the introduction of a more scientific methodology in the study of literature, looking for methodological and theoretical insights into cognitive science and evolutionary psychology.

What is more interesting for my thesis is that, not surprisingly, the debate around the legitimacy and acceptability of the cognitive approaches in literary studies has determined a discussion on the problem of literary interpretation that has many similarities with the discourse I have proposed in this paper. Recently, a young and brilliant scholar active in the field, Marco Caracciolo, has re-opened the debate, and in doing this he has explicitly stated that "In order to contribute to cognitive science, literary scholarship has to complement-and in some cases even supplant-interpretation with a different set of goals and methods". (Caracciolo, 2016: 193).

The other scientific field where literary studies can find a theoretical framework that takes great advantage of distant reading methodology is that of cultural evolution. This field of study that as of now has no application in literary studies, aims at providing a naturalist and empirical explanation of the nature and evolution of culture, adopting widely mathematical/ statistical and computational modeling.

One of the theoretical underpinnings of cultural evolution is the adoption of the population thinking framework, taken from evolutionary biology (after Ernest Mayr interpretation of Darwin's theory) and population genetics, and its application to cultural phenomena as pointed out recently by Dan Sperber and his collaborators (Claidière et al., 2014):

Literature is part of the cultural sphere, so it can be considered a population of individual items (the texts) whose members are defined by a set of measurable features. The description of the population at a given state (synchronic, in our beloved Sausurrian terms) and its evolution (diachronic) is feasible by the way of statistical and data-driven analysis.

To conclude, I think that in order to take full advantage of the most advanced methods and analytical techniques encompassed by the label Distant reading, like text mining and machine learning, in literary and cultural studies, we need to find a proper theoretical framework that gives sense to the hypothesis experiments, data sets and explanations

we can generate. The attempt to justify and anchor this approach in the context of the traditional literary theories and methodologies has proven a limitation that undermines the interesting analytical results, and it is easily amenable to the 'so what' criticism, or ideological attacks. Maybe it is time to change the framework, and to abandon the classical hermeneutical literary studies environment.

References

Bode, K. (2018). A World of Fiction: Digital Collections and the Future of Literary History. Ann Arbor, MI: University of Michigan Press.

Caracciolo, M. (2016). Cognitive Literary Studies and the Status of Interpretation: An Attempt at Conceptual Mapping. New Literary History, 47(1): 187-207 doi:10.1353/nlh.2016.0003.

Claidière, N., Scott-Phillips, T. C. and Sperber, D. (2014). How Darwinian is cultural evolution?. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1642): 20130368 doi:10.1098/rstb.2013.0368.

Da, N. Z. (2019). The Computational Case against Computational Literary Studies. Critical Inquiry, 45(3): 601-39 doi:10.1086/702594.

English, J. F. and Underwood, T. (2016). Shifting Scales: Between Literature and Social Science. Modern Language Quarterly, 77(3): 277-95 doi:10.1215/00267929-3570612.

Fish, S. (2012). Mind Your P's and B's: The Digital Humanities and Interpretation. Opinionator, New York Times http://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/ (accessed 24 January 2012).

Fish, S. E. (1980). What is stylistics and why are they saying such terrible things about it?. Is There a Text in This Class? The Authority of Interpretive Communities. Cambridge, Mass.: Harvard University Press, pp. 68-96.

Gavin, M. (2018). Vector Semantics, William Empson, and the Study of Ambiguity. Critical Inquiry, 44(4): 641-73 doi:10.1086/698174.

Marche, S. (2012). Literature is not Data: Against Digital Humanities. Los Angeles Review of Books http://lareviewofbooks.org/essay/literature-is-not-data-against-digital-humanities#.

Moretti, F. (2000). Conjectures on World Literature. The New Left Review http://newleftreview.org/A2094.

Piper, A. (2018). Enumerations: Data and Literary Study. Chicago?; London: The University of Chicago Press.

Sperber, D. (1996). Explaining Culture: A Naturalistic Approach. Oxford: Basil Blackwell.

Underwood, T. (2017). A Genealogy of Distant Reading. DHQ, 11(2) http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (accessed 20 October 2017).

Inscriptions of Israel/ Palestine

Michael Satlow

michael_satlow@brown.edu
Brown University, United States of America

Elli Mylonas

elli_mylonas@brown.edu
Brown University, United States of America

Keywords: epidoc, inscriptions, non-roman alphabets, epigraphy
Topics: Asia, English, BCE-4th Century, 5th-14th Century, scholarly editing and editions development, analysis, and methods, text encoding and markup language creation, deployment, and analysis, History, Theology and religious studies

The "Inscriptions of Israel/Palestine" (IIP) project (www.brown.edu/iip) seeks to create a corpus of inscriptions (texts written on durable materials, other than coins) from the geographical location of present-day Israel/Palestine, that date from around the sixth century BCE to the seventh century CE. The inscriptions are in Greek, Latin, Hebrew, and Aramaic. The purpose of the project is not only to allow for access and robust (and ultimately, federated) searching but also for scholarly analyses. As one of the longest running active digital epigraphy projects (with over 4,000 inscriptions entered to date), IIP provides several use cases of working with a complex and challenging multi-lingual corpus. This abstract will focus on our data modeling and approach to Linked Open Data.

IIP was an early adopter of the Epidoc schema, a customization of TEI developed especially for those working with ancient texts preserved on durable materials, such as inscriptions and coins (Elliott, Bodard, Cayless). Many users of Epidoc are in contact with each other through epigraphy.info, and the schema is continually being modified in response to user requests. The general principle, however, is that each material object on which a text is inscribed or written is treated as a discrete XML file. IIP thus gives each inscription a unique, findable ID that also serves as its document name (e.g., ash0001.xml). Each file has an extensive teiHeader, in which we encode the

metadata, where possible using controlled vocabularies as attributes of elements, linked to authority files. This kind of robust encoding allows for the database-style, faceted indexing and searching powered by SOLR that we provide through our interface. We also include images (although we have much more work to do collecting them) and geographical information. We have a geographical interface that allows for mapping (see Figure 1).

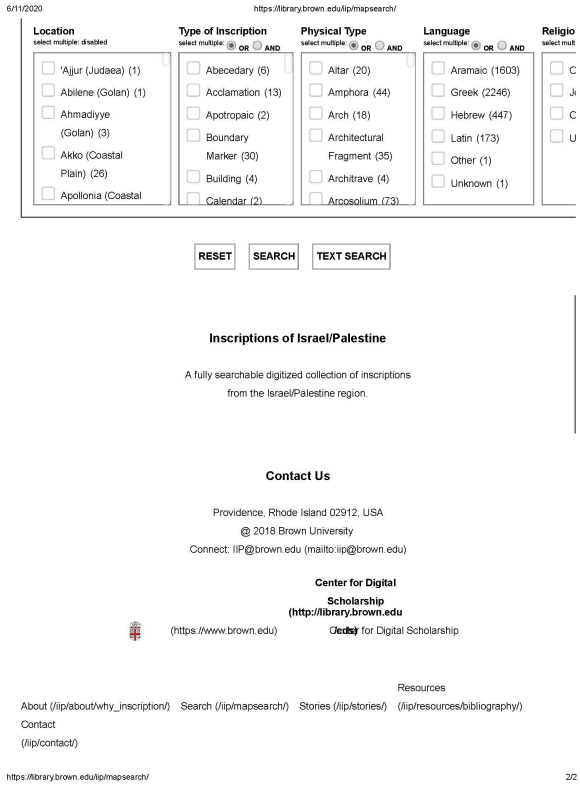


Figure 1: Screenshot of Search Page of "Inscriptions of Israel/Palestine"

Our data has always been open. All of our XML files can be seen and downloaded individually directly from our site, or downloaded in bulk from our open Github site (https://github.com/Brown-University-Library/iip-texts) or through our API (for which we give detailed instructions on the site itself). We also encode our permission license (CC BY-NC 4.0) into our files.

Since participating in the conference, "The Big Ancient Mediterranean" (BAM) we have sought to create links in our data in three ways. The primary geographical data within each inscription is linked to its corresponding Pleiades Gazetteer id (Pleiades). The primary chronological data within each inscription is linked to its corresponding PeriodO id (PeriodO). And each of the types of object upon which the inscription is written is linked to the Getty Art and Architecture Thesaurus (Getty). For example:

notBefore="0001"
notAfter="0100">First century CE
Judaea
Jerusalem
Akeldama Caves
Cave 2 Chamber B

The use of these ids allows our data to be scraped live by different projects, such as Pleiades and the Pelagios Network (Pelagios Network). The community is just now beginning to develop an ecosystem that allows for the fruitful exchange of data between sites using Linked Open Data and we hope that through this expansion the usefulness of our data will expand.

We added these links to existing data, which was a costly process. We first developed an XSLT script to extract the place, time, and object values into a spreadsheet. We then manually found and added the links, in the process having to submit new geographical and chronological values for inclusion in the other authorities (and then, in return, adding the new id numbers). We then used another XSLT script to insert the new links into the XML files. Along the way there was a great deal of checking and testing. For new data files, the additions are added at the time of creation.

One of our active projects involves the lexicographical tagging of the texts in a way that could similarly be linked and shared. This entails performing word segmentation on the existing XML files, and then assigning part of speech tags using natural language parsing. This will enable new interface features and better forms of analysis. The Global Philology Project is an exploratory project that began to lay the infrastructure for compiling and analyzing lexicographical data in many different languages across multiple sites (Global Philology). We want to further explore how our tagging of individual words could make our data - on the level of the individual words of our texts - accessible and more useful to researchers in different fields.

For a broader description of IIP and our goals, see Satlow (forthcoming) and Lembi (forthcoming). We describe our approach to bibliographical management at Lembi, Mylonas and Satlow (2016) and our approach to the FAIR principles (FAIR) in Mylonas, Lembi, Creamer, and Satlow.

References

BAM: The Big Ancient Mediterranean: https://dsps.lib.uiowa.edu/bam/2016/05/10/linking-the-big-ancient-mediterranean-conference-june-6-8-2016/

Elliott, Tom, Gabriel Bodard, Hugh Cayless et al. (2006-2019). EpiDoc: Epigraphic Documents in TEI XML: http://epidoc.sf.net.

Global Philology Project: https://www.dh.uni-leipzig.de/wo/projects/global-philology-project/

Lembi, Gaia, forthcoming, "Inscriptions of Israel/Palestine," in Sarah Schulthess, ed., Digital Biblical Studies

Lembi, Gaia, Elli Mylonas, Michael Satlow, 2016. "Bibliography in the Inscriptions of Israel/Palestine": https://tei-c.org/Vault/MembersMeetings/2016/sites/default/files/Gaia%20Lembi,%20Elli%20Mylonas,%20Michael%20Satlow%20Bibliography%20in%20the%20Inscriptions%20of%20Israel%20Palestine%20Epigraphic%20Project.pdf

Mylonas, Elli, Gaia Lembi, Andrew Creamer, Michael Satlow, forthcoming. "Archiving a TEI Project FAIRly," in TEI Conference Volume.

Pelagios Network: https://www.pelagios.org

Getty Art and Architecture Thesaurus Online: https://www.getty.edu/research/tools/vocabularies/aat/

Periodo: A gazetteer of period definitions for linking and visualizing data. https://perio.do.

Pleiades: https://pleiades.stoa.org.

Satlow, Michael L, forthcoming, "Inscriptions of Israel/Palestine," Jewish Studies Quarterly.

The GeoNewsMiner: An interactive spatial humanities tool to visualize geographical references in historical newspapers

Lorella Viola
lorella.viola@uni.lu
Utrecht University, Netherlands, The

Jaap Verheul
j.verheul@uu.nl
Utrecht University, Netherlands, The

Keywords: Spatial Humanities, diasporic newspapers, transatlantic migration, NER, machine learning
Topics: Comparative (2 or more geographical areas), Europe, English, North America, 19th Century, 20th Century, artificial intelligence and machine learning, text mining and analysis, Geography and geo-humanities, History

The GeoNewsMiner[1]: An interactive spatial humanities tool to visualize geographical references in historical newspapers[2]

The GeoNewsMiner (GNM) is an interactive tool that maps and visualizes geographical references in historical newspapers. As a use case, we analysed Italian immigrant newspapers published in the United States from 1898 to 1920, as collected in the corpus ChronicItaly (Viola 2018). Immigrant newspapers form a rich source that adds a historical dimension to the study of both the migration of the past century and the migratory experiences of migrant communities (Viola and Verheul 2019). They for instance enable researchers to compare references to the homeland and the host land (Vellon 2010; Forlenza and Thomassen 2016), thus offering an indication of the way diasporic media negotiate processes of assimilation and ethnic identification (Park 1922; Rhodes 2010; Viola and Musolff 2019, Viola and Verheul 2019), a topic that bears great relevance in the global age of satellite dishes and internet connectivity (Dhoest et al. 2012; Hickerson and Gustafson 2016; Parks 2005; Matsaganis, Katz, and Ball-Rokeach 2011; Appadurai 2008).

In order to offer new perspectives on the geographies of the past, we employed a state-of-the-art deep learning method to extract and disambiguate place names from historical newspapers. Deep learning outperforms the state-of-the-art of place name extraction and disambiguation based on static lists in gazetteers or ensembles of NER-tools (Canale, Lisena, and Troncy 2018; Won, Murrieta-Flores, and Martins 2018; Mariona Coll Ardanuy and Sporleder 2017; Maria Coll Ardanuy 2017, Yadav & Bethard 2019). The two major advantages lie in its potential for text enriching: 1) they may be based on the historical context of a historical corpus; 2) they are able to recognize toponyms in a dynamic way, for instance as as a geographical concept (Viola and Verheul 2020). For the development of the GNM, we the deep learning sequence tagging tool developed by Riedl and Padó (2018). The sequence tagging retrieved 1,369 unique locations which occurred 214,110 times throughout the whole corpus. Because each individual document is

timestamped, it was possible to quantify the number of references to each location was at any given time within the timeframe of ChroniclItaly, that is 1898-1920. Afterwards, locations were geocoded by using the Google API which identifies a place as it is stored in the Google Places database and in Google Maps. The tagged version of ChroniclItaly is available as an open access resource (ChroniclItaly 2.0, Viola 2019).

Finally, to visualise and explore the data, we developed the GNM App (Figure 1). Unique to this tool is the possibility to aggregate the data according to a wide range of parameters (time; newspaper's title; least/most mentioned places; absolute or relative frequency; aggregation on national, regional or city level). It is also possible to overlay historical maps that show the borders of selected years (1880, 1914, 1920, 1994), and download and share the data/results (Figure 2). This offers users the possibility to analyse the results in an intuitive, interactive, and reproduceable way as well as providing great flexibility to researchers working in spatial humanities, particularly from a historical perspective.

One potential application of GNM is for example the possibility to reconstruct the "geographical agenda" of historical newspapers by analysing the changing geographical bias of the press, an issue urgent to fields such as media studies, cultural history and international relations (McCombs 2014; Craine 2014; Reese and Lee 2012; Wanta, Golan, and Lee 2004; Gans 2004; Beaudoin and Thorson 2001; Ginneken 1998; Gitlin 2003). As a preliminary data exploration, for instance, the tool shows that references to geographical locations in both Italy and the United States stay remarkably stable over the period that includes the First World War.[3] The full documentation of GNM is made available to the research community to facilitate transparency, reproducibility and replicability (Viola 2020).[4] The app has much to recommend particularly to humanities scholars who are more and more confronted with the challenge of exploring collections larger than before and in a digital format.

[1] GeoNewsMiner is a project by Lorella Viola and Jaap Verheul. This project was funded by the Utrecht University Innovation Fund for Research in IT and received support from the Research Engineering team of Utrecht University. The technical implementation was provided by Jonathan de Bruin and Casper Caandorp. The Shiny app was developed by Kees van Eijden

[2] We would like to thank DH2020 reviewers for their helpful comments

[3] GNM is available as an open access app at https://utrecht-university.shinyapps.io/GeoNewsMiner/

[4] https://github.com/lorellav/GeoNewsMiner

Web 2.0 for an Authoritative Web Site

Allison Burkette
allison.burkette@uky.edu
University of Kentucky, United States of America

William Kretzschmar
kretzsch@uga.edu
University of Georgia, United States of America

Keywords: Linguistic Atlas Project, Web 2.0, digital collaboration
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, public humanities collaborations and methods, Linguistics

As we prepare to move into the next generation of the Linguistic Atlas Project website, we now must engage in theoretical discussion about a critical issue in DH: how to democratize online information in accordance with what has become known as Web 2.0. The LAP has long-standing authority within academic discussions of the development and characteristics of the different varieties of American English. One of the goals of the LAP editors is the expansion of the use of LAP data by non-linguists. In order to reach a new, wider audience we plan to create a "Teaching and Sharing" extension of the LAP website, which raises the question: how do we, as sponsors of the LAP website, negotiate authority with participation; in short, how do we let people participate without letting go of the authoritative nature of the website? This presentation addresses this question along with some possible answers.

NER on Ancient Greek texts with minimal annotation

Chiara Palladino
chiara.palladino@furman.edu
Furman University, Classics Department

Farimah Karimi
bmathiak@uni-koeln.de

GESIS Leibniz Institute for the Social Sciences

Brigitte Mathiak
University of Cologne, Institute of Digital Humanities

Keywords: Named Entity Recognition, Herodot, Conditional Random Fields
Topics: Europe, English, BCE-4th Century, natural language processing, text mining and analysis, Computer science, Philology

This paper presents the results in the adaptation of a new workflow of Named Entity Recognition and classification applied to primary sources in Ancient Greek. We used a model of language-independent data extraction and pattern discovery based on machine learning algorithms, which allowed the extraction of a dataset of automatically classified place-names and ethnonyms starting from a small manually annotated dataset. The idea is that we should be able to train the machine to recognize an entity from recurring elements in the context, without providing a long annotated training dataset in advance, working on the assumption that premodern textual sources display a recognized systematicity in their linguistic encoding of space, which provides a test-case for automatic and semi-automatic methods of pattern discovery and extraction.

THE READING EXPERIENCES ONTOLOGY (REO): REUSING AND EXTENDING CIDOC CRM

Francois Vignale
francois.vignale@univ-lemans.fr
Le Mans Université, France

Alessio Antonini
Open University, UK

Guillaume Gravier
CNRS, France

Keywords: Ontology management, CIDOC-CRM, digital heritage, history of reading
Topics: Global, Europe, French, 19th Century, 20th Century, Contemporary, data modeling, linked (open) dataBook and print history, Cultural studies

This paper aims to present the development strategy of the ontology proposed in the READ-IT project (https://readit-project.eu) and the contributions it makes to the conceptual description of reading experiences and, more broadly, to the description of intangible heritage and other "experiential" phenomena. Its development relied on a data-driven approach with the active participation of a representative panel of reading experts from HSS disciplines. The process was iterative to converge towards a consensus balancing ICT requirements and HSS scholar needs. In this regard, we will focus on the general framework and the way in which both alignments of CIDOC CRM (and some of its extensions) with READ-IT's data model and creations of classes have been carried out, as well as the benefits derived from a pragmatic and cost-efficient approach, allowing us to offer REO as an extension of CIDOC CRM, which will guarantee its reusability, improvement and maintenance over time.

From Public Humanities to Social Remembering: Big Data and the Digital Redlining of Women in Country Music Culture

Jada Emily Watson
jwatso4@uottawa.ca
University of Ottawa, Canada

Keywords: social remembering, country music, gender representation, big data, digital redlining
Topics: English, North America, Contemporary, digital biography, personography, and prosopography, public humanities collaborations and methods, Feminist studies, Musicology

Theories of social remembering (Misztal 2003; Strong 2011) and digital redlining (Noble 2018) offer a critical framework for considering the credibility of big data within cultures that disadvantage and systematically ignore women. Reflecting on results of a data-driven analysis of Mediabase's country airplay

reports from 2000 to 2018, this paper considers the role of data in the process of shaping country music culture, and reframes our understanding of these reports as an instrument that systematically "remembers" some artists, while "casting away" or "forgetting" others. Over the course of this period, the number of songs by women played on country format radio declined 41.3% (Fig. 1). These reports map the evolving terrain of country music's cultural space and have resulted in a system that pushes women to the margins: songs by women are played infrequently on country format radio (Fig. 2), with the majority of their airplay occurring in the overnights (Fig. 3) (Watson 2019a/b). As a result, songs by women are charting in declining numbers, peaking in the bottom positions of the weekly charts, and barely heard in radio's peak daytime hours. Such practices impact black women more significantly (Watson 2020). In this way, the reports reveal a digital redlining of women, wherein programming practices are perpetuating inequalities by refusing high traffic times of day to already marginalized artists. More critically, this paper addresses the challenges of critiquing social remembering through public scholarship and reflects (as Marcia Chatelain [2016] does in her work) on my experiences of thinking and working in public digital spaces.

Gender has been a central dynamic of country music history and culture (Pecknold & McCusker 2016), wherein masculinity and femininity are invoked to define class boundaries, cultural tastes, institutional hierarchies, performance styles, and the evolution socially prescribed roles. With strong ties to conservatism and religion, country's first female artists often appeared on stage with their husband or male family members, a constant reassurance for record-buyers that the social order in which females performed familial roles and habits of constancy and tradition endured in the genre (McCusker 2017). Following WWII, as female artists began taking a more prominent role on stage and behind the scenes, cultural institutions actively sought to censor lyrics in women's songs if they were too "suggestive", aggressive, or politically charged- a trend that has continued throughout the genre's history (Bufwack & Oermann 2004; Keel 2004; Watson & Burns 2010). Behind the scenes, the country music industry has employed a strict quota system for female artists on radio playlists and label rosters (Penuell 2015), which has, in turn, limited their opportunities for participating within the mainstream of the industry as performer and songwriters.

Adopting methods for data-driven studies of popular music charts (Wells 2001; Lafrance et al 2011) and influenced by the concept of prosopography (Keats-Rohan 2007; Crompton & Schwartz 2018), this project has developed an approach for collecting and organizing music industry data in order to study how the biography of individuals shapes and is shaped by the genre's cultural constructs. In order to address complex socio-cultural issues of equity and diversity in country music, it has developed a comprehensive dataset of all of the singles played on country format radio between 2000 and 2018, enhanced with biographic information about the lead and featured artists involved in performing the recorded tracks played on country radio to facilitate a vast range of queries about programming practices and their impact on weekly charts. In so doing, this project deconstructs the gender politics that have governed the genre and shows how big data has created and perpetuated gender inequalities and contributed to the continued marginalization and "forgetting" of female narrative voices within country music culture.

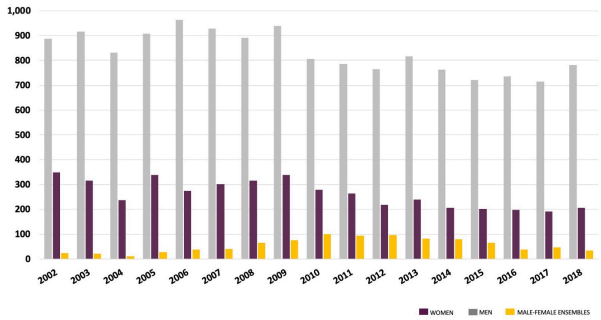


Figure 1. Distribution of unique songs by men, women and male-female ensembles played on country format radio between 2002 and 2018.

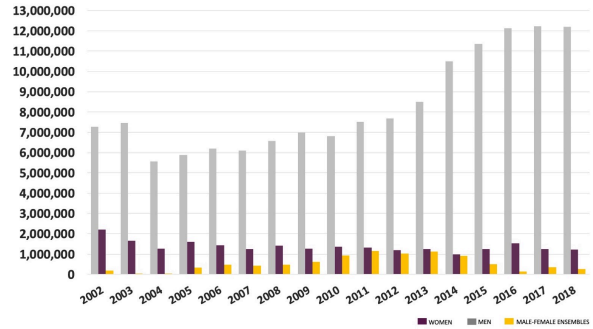


Figure 2. Distribution of spins for songs by men, women and male-female ensembles between 2002 and 2018 reveals a 40.2% increase in spins for male artists against a 44.8% decline in spins for songs by women.

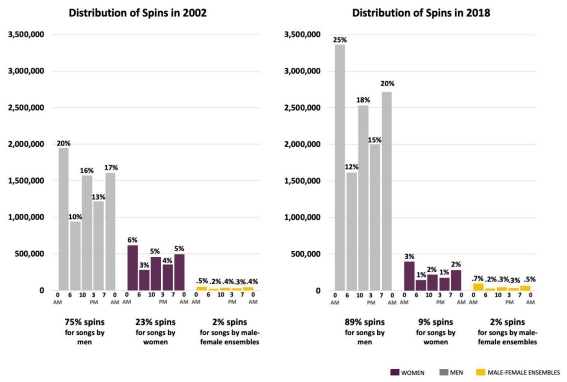


Figure 3. Distribution of spins for songs by men, women and male-female ensembles across the five dayparts in country format radio programming in 2002 (left) and 2018 (right).

Appendix A

Bibliography

Bufwack, Mary A., and Robert K. Oermann. 2003. Finding Her Voice: Women in Country Music, 1800-2000. Nashville: Vanderbilt University Press; Country Music Fouccation Press.

Chatelain, Marcia. 2018. "Is Twitter Any Place for a [Black Academic] Lady?" In Bodies of Information: Intersectional Feminism and Digital Humanities, eds Elizabeth Losh and Jacqueline Wernimont, 173-184. Minneapolis: University of Minnesota Press.

Crompton, Constance, and Michelle Schwartz. 2018. "Remaking HIstory: Lesbian Feminist Historical Methods in the Digital Humanities." In Bodies of Information: Intersectional Feminism and Digital Humanities, eds Elizabeth Losh and Jacqueline Wernimont, 131-56. Minneapolis: University of Minnesota Press.

Keel, Beverly. 2004. "Between Riot Grrrl and Quiet Girl." In A Boy Named Sue: Gender and Country Music, eds Kristine McCusker and Diane Pecknold, 155-77. Jackson: University of Mississippi Press.

Keats-Rohan, Katherine S.B.. 2007. Prosopography Approaches and Applications: A Handbook. Oxford: Oxford University Press.

Lafrance, Marc, et al. "Gender and the Billboard Top 40 Charts between 1997 and 2007." Popular Music & Society 35/5: 557-70.

McCusker, Kristine M. 2017. "Gendered Stages: Country Music, Authenticity, and the Performance of Gender." In The Oxford Handbook to Country Music, ed Travis D. Stimeling, 355-74. Oxford: Oxford University Press.

Misztal, Barbra A. 2003. Theories of Social Remembering. Philadelphia: Open University Press.

Noble, Safiya Umoja. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. New York: New York University Press.

Pecknold, Diane, and Kristine M. McCusker, eds. Country Boys and Redneck Women: New Essays in Gender and Country Music. Jackson: University of Mississippi Press, 2016.

Penuell, Russ. 2015. "On Music and Scheduling." Country Aircheck, 449: 1, 8, https://www.countryaircheck.com/pdfs/current052615.pdf.

Strong, Catherine. 2011. "Grunge, Riot Grrrl and the Forgetting of Women in Popular culture." The Journal of Popular Music Culture 44/2: 398-416.

Watson, Jada. 2019a. "Gender Representation on Country Format Radio: A Study of Published Reports from 2000 to 2018." Report prepared in consultation with WOMAN Nashville, https://bit.ly/gender-country-radio.

Watson, Jada. 2019b. "Gender Representation on Country Format Radio: A Study of Spins Across Dayparts (2002-2018)." Report prepared in consultation with WOMAN Nashville, https://bit.ly/gender-country-radio-tod-airplay.

Watson, Jada. 2020. "Inequality on Country Radio: 2019 in Review." Report prepared in partnership with CMT's EqualPlay Campaign, https://bit.ly/gender-country-radio-equalplay.

Watson, Jada, and Lori Burns. 2010. "Resisting Exile and Asserting Musical Voice: The Dixie Chicks Are 'Not Ready to Make Nice'." Popular Music 29/4: 325-50.

Wells, Allan. 2001. "Nationality, Race and Gender on the American Pop Charts: what Happened in the 90s?" Popular Music & Society 25: 221-31.

Project Twitter: Literature Scraping, Analyzing, and Archiving Twitter Data in Literary Research

Christian Howard-Sukhil
cfh008@bucknell.edu
Bucknell University, United States of America

Keywords: social media, data curation, data preservation, globality
Topics: Global, English, Contemporary, digital archiving, social media analysis and methods, Literary studies, Media studies

Project Twitter Literature (TwitLit), seeks to address a growing gap in the literary-historical record[1] by establishing a consistent, rigorous, and ethical method for scraping and cleaning up Twitter data for the use of humanities scholars. In particular, my project explores the growing community of amateur writers who are using Twitter as a means of publication and dissemination for their literary output. There are three parts to my project: the research findings related to the global literary community on Twitter, the tools and resources developed as part of the project and made openly available to other scholars, and partnership with a university library to ensure the long-term preservation of the collected data.

The data that I have collected shows that social media is altering literary practices by providing a space for amateur writers to publish, disseminate, and receive feedback from a global community of writers. Preliminary figures put the number of active Anglophone writers using Twitter as a publication platform for their literary output at over 1 million users per year since 2015, and writers working in non-English languages on Twitter raise these numbers even higher. This practice is changing how literature is produced, published, and shared. Readerships too are changing, for rather than being tied to print subscriptions or access to physical books, audiences of social media literature are based on online communities and tied to the costs of physical devices and internet access.

My presentation will showcase these research findings in order to highlight the importance and necessity of social media archival work. In so doing, I will discuss how I collected the data using a Python script (co-developed by myself and several other scholars), challenges of cleaning up and visualizing this data (using ArcGIS and tools developed by Documenting the Now), and ethical best-practices for using social media data in research. Information relating to this process - including detailed instructions, the Python scripts used to collect Twitter data, and a list of resources - are free and openly accessible on the project's website (www.twit-lit.com) and GitHub repository (https://github.com/TwitLit/TwitLitSource). Other scholars are invited to use these scripts and other resources to collect their own social media data.

Additionally, my project has attempted to plan for the long-term preservation of the over eight million tweets that I have collected. This preservation has been made difficult by Twitter's strict Developer Policy and Agreement, which prevents individuals from keeping or disseminating large data sets for more than 30 days. The only exception to this policy is made on behalf of academic institutions, which may store Twitter data for unlimited amounts of time on behalf of academic research.[2] The Project TwitLit project thus presents best-practices for establishing a working relationship with university libraries for storing and disseminating Twitter data in a way that is both in accord with Twitter's legal restrictions and responsive to the needs of scholars. In short, Project TwitLit provides a case-study of a growing community on Twitter while simultaneously developing a set of tools and guidelines for other scholars seeking to engage in similar work.

[1] In December 2017, the Library of Congress, which began archiving Twitter in 2010, announced that it would no longer collect all Tweets; instead, Tweets produced after December 2017 would only be collected on a selective basis. There are no other ongoing, systematic efforts to collect and preserve this digital material. See Library of Congress, "Update on the Twitter Archive at the Library of Congress" (December 2017).

[2] For more information related to the challenges of collecting and storing Twitter data, please see Christian Howard, "Studying and Preserving the Global Networks of Twitter Literature," in Post-45.

"Beyond the logic of commensurability a cultural analysis of media artworks and digital media in information capitalism"

Kevin Day
kevin.t.day@gmail.com
University of British Columbia

Keywords: media art, philosophy of technology, media studies, information capitalism, big data
Topics: Global, English, Contemporary, digital art production and analysis, Art history, Media studies

The proposed presentation will examine the pedagogical potential of media artworks that interrogate the big data economy within contemporary information society. The study begins by establishing the socio-political landscape within which it is situated, one that recognizes the pervasive utopic myth and exploitative algorithmic activities of informatics, and

asserts that media art needs to address digital media by examining the underpinning logic of information within the wider landscape of information capitalism. Guided by a framework that pulls together theories of media and art, the paper argues that media art has the capacity to subvert normalized and entrenched ways of knowing through its potential to foster ways of 'knowing differently' in relation to information and communication technology. To substantiate the argument, the paper will position information as an epistemic model through which one comes to make sense of the world - and precisely that which visual/media art should tackle and question.

The Plant Humanities Workbench Using Linked Open Data to Discover Early Modern Plant History

Nathan Kelber
NKELBER@GMAIL.COM
ITHAKA

Keywords: plants, linked open data, early modern, book history
Topics: Comparative (2 or more geographical areas), Europe, English, North America, 15th-17th Century, Contemporary, linked (open) data, software development, systems, analysis and methods, Book and print history, Environmental, ocean, and waterway studies

In 2017 with the support of the Mellon Foundation, the experimental wing of JSTOR known as JSTOR Labs began a three year partnership with Dumbarton Oaks, a Harvard-owned research library and collection in Washington DC, to create a new digital humanities tool focused on the study of plants from a humanities perspective. The Plant Humanities Workbench will enable new research by intergenerational teams of students, advanced researchers, and professionals. Drawing on the technical and design expertise of the JSTOR Labs team, the digital tool will be interactive, iterative, and scalable. This presentation will demonstrate the tool after its first year of development, summarizing key findings in the project that may be relevant to a variety of digital humanists.

Bitter Wind: Adapting Greek Tragedy for Spatial Computing

Elizabeth Hunter
ebh@sfsu.edu
San Francisco State University

Keywords: spatial computing, theatre studies, Greek tragedy, mixed reality, interactivity
Topics: EnglishNorth AmericaBCE-4th Century, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), virtual and augmented reality creation, systems, and analysis, Literary studies, Performance Studies: Dance, Theatre

This presentation will report on the completed research project Bitter Wind, my adaptation of an ancient Greek tragedy for Microsoft's spatial computing HoloLens headset. The presentation will demonstrate how emerging embodied technologies like spatial computing (1) offer new possibilities for qualitative interpretation and (2) promote greater theatre studies engagement with digital humanities. In addition to outlining Bitter Wind's technological elements, the presentation will explain how I integrated the affordances of spatial computing with theories of audience participation to mount a new dramaturgical and historiographic analysis of a canonical source. This presentation will also argue that a stronger theatre studies presence in DH has become urgent, as DH inquiry begins to expand from two-dimensional screens to the immersive, 360° environment created by embodied technologies like augmented/mixed/virtual reality, wearables, and the Internet of Things.

DUCAT Passage/ Translation Alignment with the CITE Architecture

Christopher William Blackwell
cwblackwell@gmail.com
Furman University, United States of America

Chiara Palladino

mackense.greico@furman.edu
Furman University, United States of America

MacKense Greico
Furman University, United States of America

Allie Bolton
Furman University, United States of America

Keywords: citation, translation, prose, poetry, alignment
Topics: Comparative (2 or more geographical areas), Europe, English, North America, Contemporary, linked (open) data, semantic analysis, Literary studies, Translation studies

DUCAT is a tool that uses the CITE Architecture, allowing alignments among texts. It is a zero-infrastructure HTML app that exports data in a plain-text CEX format. It allows two kinds of alignments among any number of texts, in any language:

Citation Alignment A poem, cited line-by-line, is translated into prose, cited by paragraph. A paragraph of the translation might correspond to several lines of poetry; a range of lines of poetry might overlap with seveal paragraphs of the prose translation.

Translation Alignment A text and its translation will not align word-by-word, but will inevitably consist of one-to-one, one-to-zero, many-to-one, one-to-many, many-to-many, many-to-zero, or zero-to-many alignments. In the case of "many" the words may be discontiguous. When we align more than one translation with an original text, these problems multiply.

While DUCAT is a standalone tool, the data it produces is generic, plain-text, and can be re-imported, shared, and reused.

MaDiH (مديح) Mapping Digital Cultural Heritage in Jordan

James Smithies
james.smithies@kcl.ac.uk
King's College London

Fadi Balaawi
Hashemite University

Pascal Flohr
University of Oxford

Shaher Rababeh
Hashemite University

Sahar Idwan
Hashemite University

Carol Palmer
Council for British Research in the Levant

Shatha Mubaideen
Council for British Research in the Levant

Alessandra Esposito
King's College London

Arianna Ciula
King's College London

Keywords: Jordan, archaeology, infrastructure, data, analysis
Topics: Asia, English, BCE-4th Century, 5th-14th Century, 15th-17th Century, data publishing projects, systems, and methods, database creation, management, and analysis, Archaeology, Library & information science

MaDiH (مديح): Mapping Digital Cultural Heritage in Jordan, is a collaborative project between King's Digital Lab (KDL), the Hashemite University, the Council for British Research in the Levant (CBRL), the Department of Antiquities of Jordan, the Jordanian Open Source Association, and the Endangered Archaeology in the Middle East and North Africa (EAMENA) project. It is scheduled to run for two years, from February 2019 - February 2021. The project will contribute to the long-term sustainable development of Jordan's digital cultural heritage, identifying key systems, datasets, standards, and policies, and aligning them to government digital infrastructure capabilities and strategies. Defining a robust technical and operational architecture for digital cultural heritage will assist the Department of Antiquities in their planning processes, help product development teams develop their systems, facilitate the aggregation of valuable datasets held in disparate repositories, and ensure data generated

from research activity is properly stored and widely accessible.

Modelling Consultation Workflows for Research Data Management in the Humanities

Patrick Helling
patrick.helling@uni-koeln.de
University of Cologne, Germany

Keywords: Research Data, Research Data Management, RDM, Data Center, Workflow
Topics: Europe, English, Contemporary, digital research infrastructures development and analysis, sustainable procedures, systems, and methods, Library & information science

Abstract

Research data management (RDM) is becoming more important for researchers and research institutions like universities and non-university centers. Due to the heterogeneity of methods in the Humanities, RDM in the Humanities is very complex. Several data centers and contact points have been established to handle this complexity, support researchers and conduct active RDM. Nevertheless, there are few best practices and policies on how RDM is done in a sustainable and goal-oriented way. In this paper I present a consultation-protocol-template we developed at the Data Center for the Humanities (DCH) at the Faculty of Arts and Humanities at the University of Cologne (UoC) to structure and measure RDM in a comprehensive way. I illustrate generic RDM-consultation-workflows developed on the basis of our consultations.

1. Introduction and Background Research

data is an essential ingredient and facilitator of scientific progress across all disciplines and includes heterogenous data (Bryant, Lavoie, Malpas, 2017). The developments of the European Open Science Cloud (EOSC) and OpenAIRE, and the plans to establish a national research data infrastructure (NFDI) for the whole German research landscape (RfII, 2016) illustrate that the relevance of RDM has reached a high political level, nationally and internationally.[1] However, this is a rather new development and practical implementation has not yet reached maturity.

The Data Center for the Humanities (DCH) is the first contact point for researchers in the Humanities at the University of Cologne (UoC) on questions concerning we (Blumtritt, Helling, Mathiak et. al., 2018).[2] Through open consultation hours the DCH advise researchers at the Faculty and beyond, attend to their projects and support them as far as RDM is concerned. In addition, we run the Language Archive Cologne (LAC), a repository for audio-visual data, we do active data management and curate, archive and publish research data.[3] In this paper, I seek to further the professionalization of RDM, by looking closer at the consultation process and the experiences made with it at our institution in 6 years of consultation practice.

2. RDM consultation protocols

In order to identify goal-oriented solutions for RDM-needs and as a basis for active RDM, we developed a consultation-workflow (see Figure 1). Through semi-structured interviews and structured consulting protocols, we are offering sustainable RDM-services and solutions (Helling, Blumtritt and Mathiak, 2018). With more than 70 consulting protocols, the DCH has a wealth of knowledge.

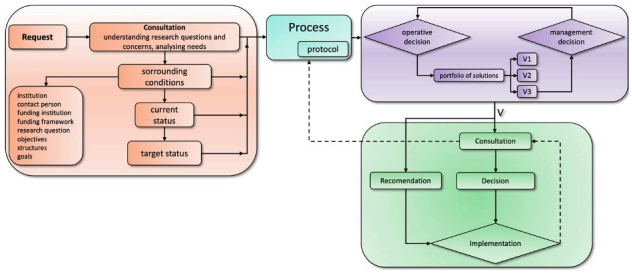


Figure 1: The core consulting-workflow at the DCH.

On the basis of a comparative analysis of the structured consulting protocols, I was able to identify the core information required for active RDM. The information can be subsumed under the categories (red) context information of a process, (yellow) RDM request-categories with a concrete description of the categories in the context of a consultation and (green) recommendations (see Figure 2).

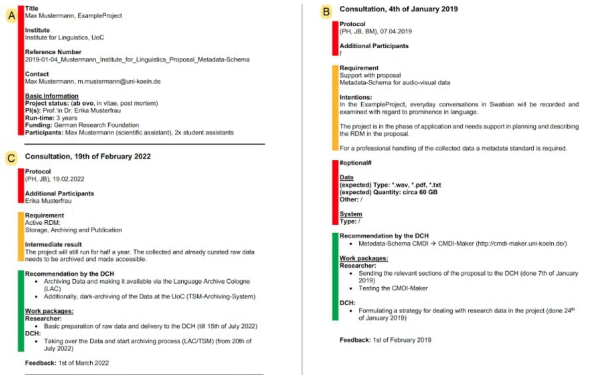


Figure 2: anonymous example of the consulting protocol template: (A) consists of meta information about a process, (B) and (C) are both consulting protocols of a process.

3. Modelling of RDM-consultation workflows

In 2019, I identified 28 demand-categories in 36 consultations. I started modelling generic workflows for each demand-category fitting with every consultation, where the demand was given with the help of the Business Process Model and Notation (BPMN).[4] I will only illustrate on two examples in this abstract, but plan to show more in the presentation.

3.1. Support with a proposal

Many funding institutions ask for statements on how researchers will deal with the digital research data representing the output of an applied project. If a consultation is needed in such a case, we conduct an interview with the researchers and try to receive as much information as possible about the project, which we translate into our consultation-protocol-template.

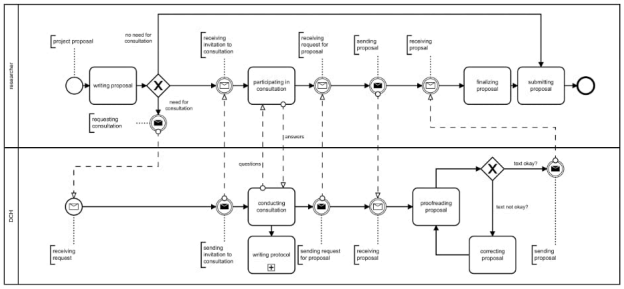


Figure 3: Consultation workflow for support with a proposal.

Based on the information in the protocol, we then follow the steps as detailed in Fig. 3, always taking the context information into account. Based on similar cases, we suggest a text to describe possible strategies for data management and define fundamental steps of data handling, archiving and publishing.These formulations get integrated into the overall application by the researchers and thus become part of the proposal.

3.2. Data archiving

Another common request is the archiving (see Figure 4) of data at the end of a project. After the generic steps of a workflow described already, we start to search for a fitting repository with respect to the conditions of the actual data and the whole project.

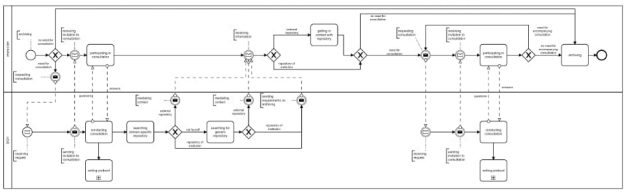


Figure 4: Consultation workflow for data archiving.

We developed a system of priorities that we work through step by step: at first, we check if we can identify a (1) domain-specific repository that fits with the given needs and conditions. In our case, this can be the LAC, or an external repository. If there is no solution, we check if there is a fitting (2) generic repository. In our case, this can only be an external repository, but, in the context of another data center, it could be a generic repository driven by the data center.

In any case the results of the investigation need to be passed to the researchers: either we mediate a contact to an external repository, or we present information about the conditions of archiving data in the LAC.

Conclusion

The formal description of our workflows optimizes both consultations and active RDM. The workflows follow acknowledged standards as well as know-how and experiences made by data managers since 2013. Thanks to these procedures, we are able to deal with the diversity of RDM-requirements in the Humanities in an effective and standardized way and we can handle different consultations with the same or a similar request equally. This makes RDM measurable and comprehensible. Because of the abstract level of modelling, the workflows are generic and can be adapted in different contexts.In my talk, I will present more workflow models and in a more detailed way. Contextually, I will define RDM-requirements and -services based on active data management in a comprehensive way and show synergies between different needs and service structures. This will lead me to a plea for the necessity of measurable RDM-structures based on RDM done in expert centers by data manager, as opposed to the common approach of building and establishing contact points and training a mediating data manager based on recommendations of political and scientific committees and umbrella

organizations that are not closely related to daily research and RDM.

Footnotes

[1] European Open Science Cloud (EOSC), Online: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud (last request: 2nd October 2019); OpenAIRE, Online: https://www.openaire.eu/ (last request: 2nd October 2019); National Research Data Infrastructure (NFDI), Online: https://www.dfg.de/foerderung/programme/nfdi/ (last request: 2nd October 2019).

[2] Data Center for the Humanities (DCH), Online: http://dch.phil-fak.uni-koeln.de/index.html (last request: 2nd October 2019).

[3] Language Archive Cologne (LAC), Online: https://lac.uni-koeln.de/ (last request: 2nd October 2019).

[4] cf. Silvere, Bruce. BPMN method and style. Cody-Cassidy Press, Aptos, CA 2009.

Bibliography

Blumtritt, J., Helling, P., Mathiak, B. et. al. (2018): Forschungsdatenmanagement in den Geisteswissenschaften an der Universität zu Köln, in: o-bib, Das offene Bibliotheksjournal, P. 104-117. DOI: https://doi.org/10.5282/o-bib/2018H3S104-117.

Bryant, R., Lavoie, B., Malpas, C. (2017): A Tour of the Research Data Management (RDM) Service Space. The Realities of Research Data Management, Part 1. Dublin, Ohio: OCLC Research. DOI : https://doi.org/10.25333/C3PG8J.

Helling, P., Blumtritt, J., Mathiak, B. (2018): Der Beratungs-workflow des Data Center for the Humanities an der Universität zu Köln, in: o-bib, Das offene Bibliotheksjournal, P. 248-261 . DOI: https://doi.org/10.5282/o-bib/2018H4S248-261.

RfII (2016): Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, 160 S. Online: http://www.rfii.de/?p=1998 (last request: 2nd of October 2019).

Identifying relations between characters in Afrikaans, Tshivenda, and Xitsonga book

Menno van Zaanen

South African Centre for Digital Language Resources, South Africa

Benito Trollip

South African Centre for Digital Language Resources, South Africa

Phathuthsedzo Ramukhadi

South African Centre for Digital Language Resources, South Africa

Respect Mlambo

respect.mlambo@nwu.ac.za
South African Centre for Digital Language Resources, South Africa

Keywords: African languages, literary analysis, named entity recognition, network analysis
Topics: Africa, English, Contemporary, electronic literature production and analysis, natural language processing, Literary studies

The usefulness of computational linguistic tools, such as named entity recognition (NER) systems, in linguistic or literary studies of under-resourced languages is an area that is still relatively unexplored. We applied NER systems to one Afrikaans novel and two scanned dramas, one in Tshivenda and one in Xitsonga. Personal relations are identified through character name co-occurence in sentences and these relationships are visualized using Gephi, following the approach by Van de Ven et al. (2018). The research identified several practical problems: low quality OCR, low quality NER, limited amounts of NE and language specific issues.

Never Alone, Never Finished: Defining Never Alone's World Games Genre as Ethical Alternative to Empathy Games

Jenn Olive

jolive1@gsu.edu
Georgia State University, United States of America

Keywords: Never Alone, world games, survivance, empathy games

Topics: Global, English, North America, Contemporary, electronic literature production and analysis, mixed-media analysis, Games studies, Literary studies

In this paper, I will investigate the genre of world games through the game that launched the genre into the spotlight, Never Alone. This debut title for the first indigenous-owned game company in the United States, Upper One Games, uses this genre to categorize itself within the rest of game experiences, which is significant given the genre's lack of use in games at large as well as what it signifies in its difference from established genres. Based on my investigation, I will argue that the potential importance of the genre as defined through Never Alone comes from its opportunity to answer to the ethical concerns of the empathy games genre.

Intersections of the Cultural Kind: Public Digital Humanities and the Museum

Patricia Bentley
bentley.pbentley@gmail.com
York University

Keywords: museum, data analysis, methodology
Topics: Asia, English, North America, 19th Century, 20th Century, Contemporary, cultural analytics, data modeling, Galleries and museum studies

This presentation examines a recent research intervention that used data mining and data mapping and visualization methods to assess how visitors in a museum of Islamic art were making sense of their encounters with patterned works of art in the galleries.

Using computer vision to identify graphic elements in Picasso's poetry

Enrique Mallen
Sam Houston State University, United States of America

Luis Meneses
ldmm@uvic.ca
University of Victoria, Canada

Keywords: computer vision, graphic elements, poetry
Topics: Europe, English, North America, 20th Century, image processing and analysis, manuscripts description, representation, and analysis, Art history, Humanities computing

Pablo Picasso started writing in 1935. The onset of this new endeavor is said to have coincided with a devastating marital crisis-a financially risky divorce to be exact-decreasing substantially his pictorial output. Writing now became his alternative outlet. Not surprisingly, Picasso's poetry is quite visual. While other poets in the past have added graphic elements to enhance their poems, their use in Picasso is of special importance as they are an essential component of his writing. Furthermore, the method he used in his poetic compositions is essentially combinatorial in nature, with words co-occurring often in unexpected ways, this relates to his collage technique during Cubism. Examples of these features are shown in figures 1 and 2.

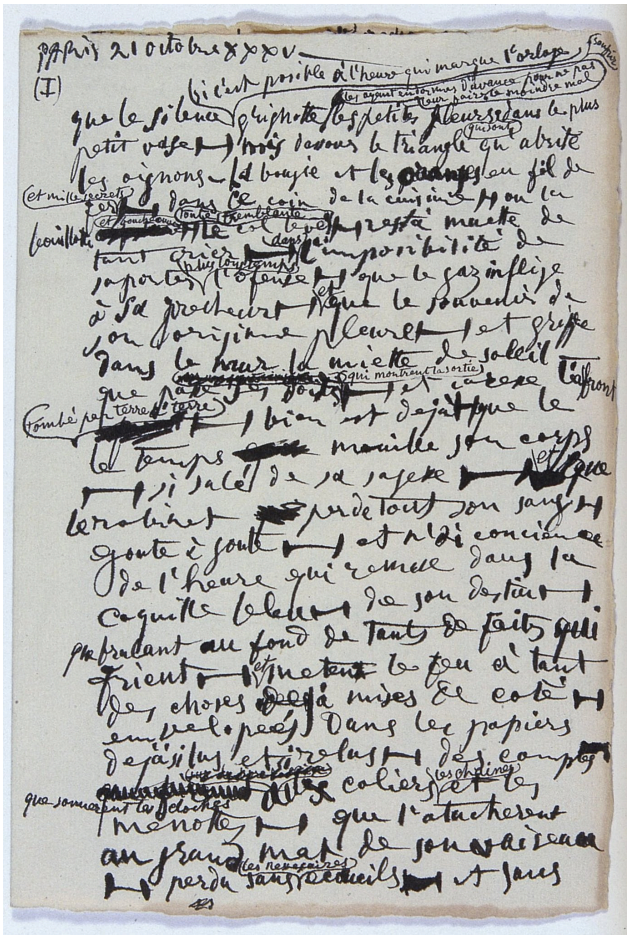


Figure 1: P. Picasso, OPP35-138, 1935

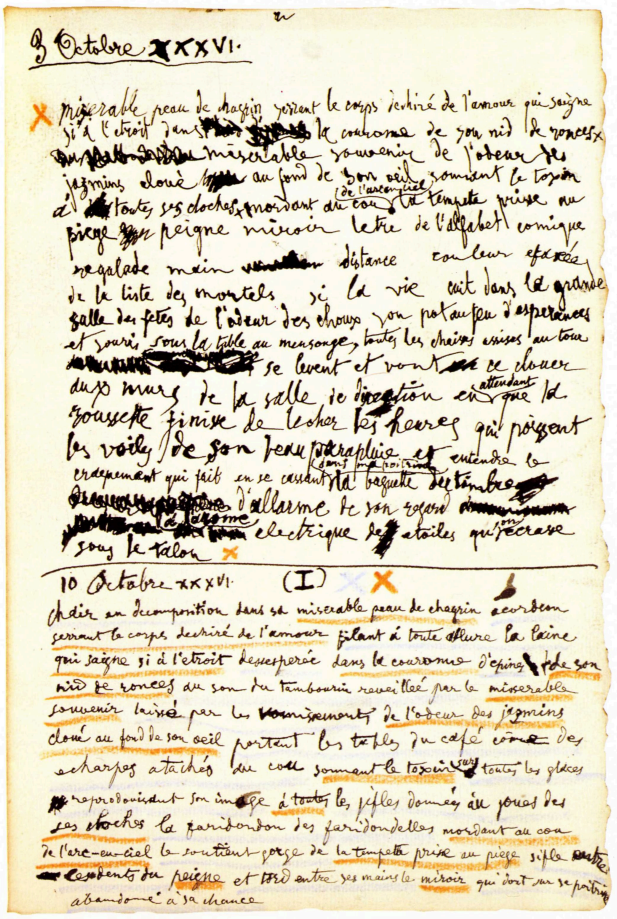


Figure 2: P. Picasso, OPP36-138, 1936.

Among the graphic elements Picasso includes in his poems are dashes, brackets, arrows, lines, blots, etc. The combinatorial nature of his writing highlights the interconnections between words and graphic elements (Mallen 2003). This combination of unique graphic and verbal components in Picasso's poetry leads us to view the poems as a set distinct layers of text and images that form the layout of a document-making it quite hard to define different boundaries while providing an important contribution to the interpretation of the layers of text. Additionally, Picasso's poems are often revisited multiple times, and he kept a record of all the changes from one state to another at various moments of their creation. Each layer with new additions and deletions is carefully dated by the author, indicating the importance that each of them had for him. The "final" published copy is artificial in the sense that the printed version has been subjected to a frozen linearized transcription (often carried out by Picasso's secretary, Jaime Sabartés) and fails to portray the information in the additional layers that can be found in the original manuscripts which fortunately have been preserved.

We previously concluded that there is a close correlation between graphic elements and the verbal context in which they occur (Meneses and Mallen 2017) and offered a solution to encode graphic features and stratified text in machine-readable form (Mallen and Meneses 2019). However, identifying and encoding the graphical elements in the corpus of poems can be a laborious and intensive process. In this paper we will present a new approach that uses OpenCV, a library of programming functions aimed at real-time computer vision (OpenCV 2019), to identify the graphic elements in Picasso's poetry. More specifically, we will present our findings on training the models, and discerning between false positives and negatives. This analysis will get us closer to understanding how the presence of graphic elements in a specific line of a poem affects the interpretation of the word string in that line and of the poem as a whole. Our final goal is to provide a systematic encoding of both the text and the graphic elements so that their mutually dependent interpretation may be properly evaluated.

References

Mallen, Enrique. 2003. The Visual Grammar of Pablo Picasso. New York: Peter Lang.

Mallen, Enrique, and Luis Meneses. 2019. "Visual Text: Encoding Challenges in Picasso's Poetry." Journal of Fine Arts 2 (2): 31-37.

Meneses, Luis, and Enrique Mallen. 2017. "Visual Text: Encoding Challenges in Picasso's Poetry." presented at the TEI 2017, Victoria, BC, Canada, November 11.

OpenCV. 2019. "OpenCV." Accessed October 11. https://opencv.org/.

Evolución y escenario actual de las Humanidades Digitales en España

Maurizio Toscano
maurizio.toscano@gmail.com
Universidad de Granada

Aroa Rabadán
Universidad Complutense de Madrid

Salvador Ros
Universidad Nacional de Educación a Distancia (UNED), Spain

Elena González-Blanco
Universidad Nacional de Educación a Distancia (UNED), Spain

Keywords: Spain, Digital Humanities, digital infrastructures
Topics: Europe, Spanish, Contemporary, database creation, management, and analysis, History of science, Humanities computing

Introducción

Las Humanidades Digitales (HD) se han convertido en un campo de interés en España, especialmente en la última década, a pesar de haber llegado más tarde que en la mayoría de los demás países europeos. De hecho, representan una tendencia destacada en la investigación, ya sea como campo de estudio que como tema de financiación preferente. Al mismo tiempo, por su novedad, están siendo objeto de escrutinio por parte de la comunidad investigadora y de las instituciones gubernamentales que financian la investigación. El objetivo del estudio ha sido identificar a los investigadores que trabajan en el campo de las HD en España y explorar su financiación, sus afiliaciones institucionales, las temáticas de investigación y los recursos digitales desarrollados.

En el pasado se han promovido iniciativas similares, que han producido mapeos centrados en el ámbito internacional o nacional, algunos de los cuales siguen disponibles en línea, con diferencias que van desde la cobertura geográfica hasta el tipo de datos mapeados (Ortega; Eunice-Gutiérrez, 2014; Romero-Frías; Del-Barrio-García, 2014). Otros estudios sobre este tema han optado por un enfoque diferente, utilizando la bibliografía u otras fuentes para identificar las etapas más relevantes en la evolución y consolidación de esta disciplina en España (Rojas-Castro, 2013; González-Blanco, 2013; Spence; González-Blanco, 2014; Baraibar-Echeverria, 2014). El presente trabajo no pretende ser una revisión histórica exhaustiva, sino ofrecer una visión complementaria y actualizada del panorama de las HD en España, tomando en consideración datos recientes y fuentes de información no explotadas anteriormente.

Materiales y métodos

A continuación se describen los resultados de la investigación sobre el estado actual de las infraestructuras de investigación digital en España, entendidas como la combinación e integración entre los recursos de información digital, las herramientas analíticas y de visualización y la comunidad activa de investigadores, colaborando a través de proyectos de investigación, financiados por el sector público o

privado. Por esta razón, hemos subdividido el objeto de la investigación en cinco entidades principales, concretamente: investigadores, proyectos, recursos, bibliografía y cursos de posgrado.

Nuestra metodología de recopilación de datos ha tenido una doble vertiente. Por un lado, hemos seleccionado manualmente información disponible en línea a partir de congresos, seminarios, convocatorias temáticas, mapas participativos, etc. Por el otro, hemos extraído información a partir de bases de datos existentes: publicaciones científicas en el campo de las HD presentes en Dialnet y en ÍnDICEs-CSIC y proyectos de investigación financiados a través de la Agencia Estatal de Investigación, seleccionados a través de una serie de palabras claves (Figura 1).

El volumen total de registros recopilados ha sido 1.359, distribuidos de la siguiente manera: 577 investigadores; 368 proyectos; 88 recursos; 9 cursos de posgrado y 8 revistas científicas. El conjunto de datos analizados se encuentra disponible en Acceso Abierto (https://doi.org/10.5281/zenodo.3893546), junto con los Jupiter Notebooks y el código Python para reproducir los análisis.

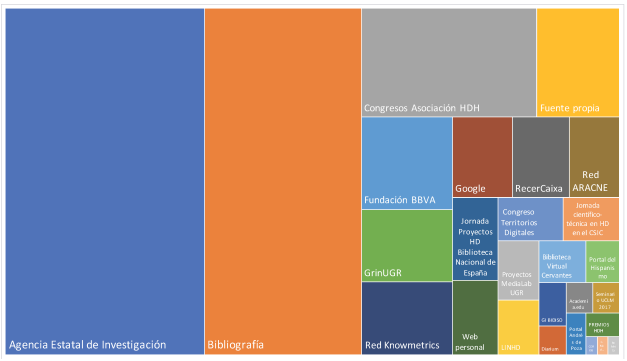


Figura 1. Fuentes de datos utilizadas en la investigación.

Resultados

- Entre los investigadores identificados, 305 son varones (52,9%) y 272 mujeres (47,1%): una proporción que, comparada con la proporción de género entre los investigadores de España en todas las disciplinas (61,2% de varones y 38,8% de mujeres) o limitada a las Humanidades (59,8% de varones y 40,2% de mujeres), refleja una presencia femenina significativamente mayor de la esperada.
- La clasificación de los investigadores por disciplinas, 19 en total, muestra una amplia variedad, con una clara prevalencia de filólogos (36%), seguidos por historiadores (16,5%) e informáticos (10,8%).
- El análisis de las conexiones entre disciplinas y temas de investigación (Figura 2) revela cinco grandes grupos disciplinarios: Historia (32% de los nodos), Filología (24%), Comunicación (20%), Ciencias de la Computación (17,3%) y Documentación (6,7%). Las Ciencias de la Computación, a pesar de ser la cuarta comunidad en términos de nodos, resulta ser la disciplina relacionada con el mayor número de temas de investigación (13,3%).

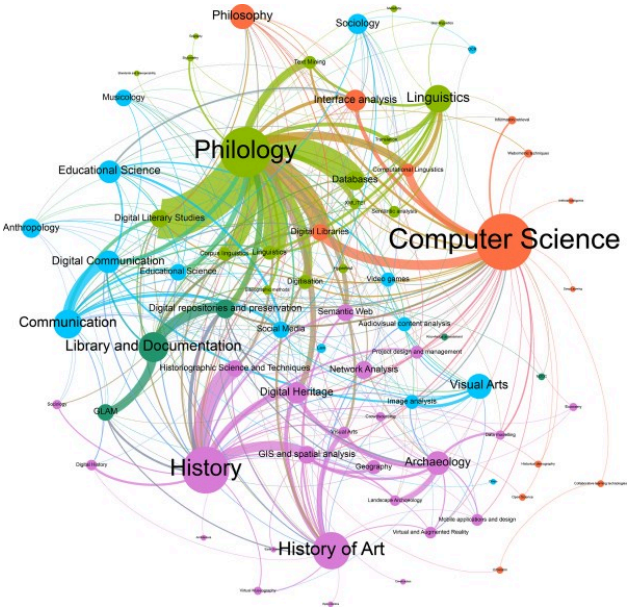


Figura 2. Análisis de redes de disciplinas y temas de investigación. El tamaño de los nodos corresponde al número total de conexiones, el color a la comunidad y el grosor de las aristas al grado de conectividad entre parejas de nodos.

- Según la afiliación institucional, la mitad de los investigadores (49,2%) pertenece a un total de nueve instituciones, mientras que la otra mitad se encuentra dispersa entre 84 diferentes centros (http:// sl. ugr.es/DHmap).
- La afiliación a departamentos (Figura 3) muestra un patrón mucho más variado en cada institución: por ejemplo, el peso relativo de la Filología es mayor en la Complutense de Madrid o en Santiago de Compostela con respecto a Granada, donde por otra parte se aprecia mucha variedad, con investigadores procedentes de 13 departamentos.

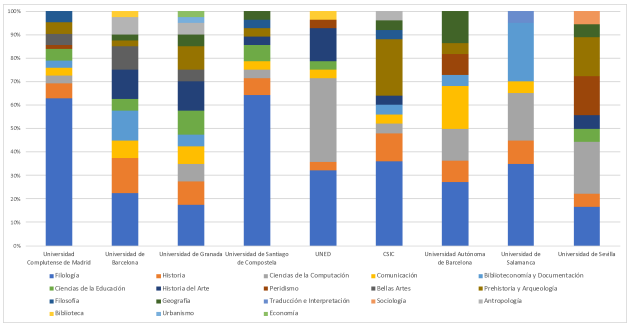


Figura 3. Peso relativo de cada departamento en los nueve centros de investigación principales. Para favorecer la comparación, se ha utilizado un histograma apilado.

- El análisis de la inversión en 337 proyectos de investigación en los últimos 25 años (Figura 4) permite avanzar una periodización en tres fases en la consolidación de as HD en España: 1993-2003; 2004-2014; 2015-2019. La financiación media por proyecto ha sido de 64.313€, pero la mediana se encuentra en 42.350€; el 5% de los proyectos ha recibido hasta 5.000€ (micro-proyectos) y el 90% menos de 100.000€.

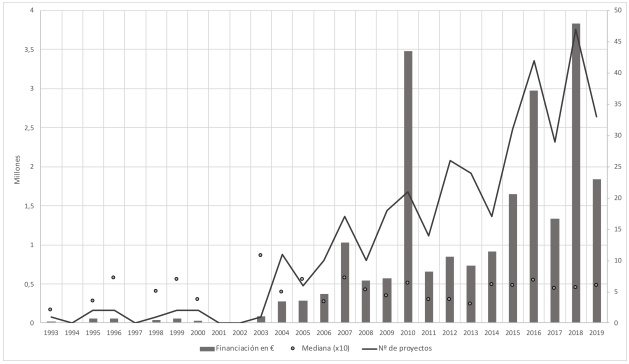


Figura 4. Total de proyectos de investigación en HD y financiación recibida durante el período 1993-2019.

- Entre las fuentes de financiación, el rol del ministerio es predominante, con 77% de las propuestas financiadas y 72% de los recursos asignados, a pesar de la gran variedad de organismos implicados, hasta 26, y de un 10% de recursos procedentes del sector privado.
- Disciplinas como la Filología, la Lingüística y la Biblioteconomía evidencian una tradición más larga que otras, como la Historia, la Arqueología y la Historia del Arte, en el desarrollo de recursos digitales para la investigación (Figura 5). Los artefactos desarrollados con mayor frecuencia (72,4%) son diferentes tipos de bases de datos (bibliotecas digitales, catálogos, repositorios, etc.), mientras que a partir de 2014 se observa un cambio hacia instrumentos más analíticos o participativos. La sostenibilidad de las plataformas en le tiempo se ha garantizado principalmente mediante la financiación en serie de proyectos del Programa Estatal de I+D+i.

	Filología	Lingüística	Historia	Arqueol.	Hist. Arte	Patrimonio	Musicol.	Document.	Comunic.	Educat.	TOTAL
Bibliotecas digitales	12						1	1			14
Bases de datos	17	2	3	2	1		1				26
Repositorios	3		2		1	4		6		1	16
Corpus	1										1
Catálogos	1	1		2	1	1					6
Crowdsourcing			1	1	1						3
Portales	4	1	2	1	1	1		1	1		12
Diccionarios		1									1
Mapas				1			1		2		4
Aplicaciones web		1			1						2
Aplicaciones móviles				2							2
TOTAL	38	6	8	9	6	6	3	8	3	1	88

Figura 5. Recursos digitales clasificados según disciplina y tipología.

Conclusiones

A modo de conclusión, podemos destacar los siguientes aspectos: (1) la mayoría de las evidencias detectadas por otros estudios se han confirmado numéricamente; (2) los análisis cuantitativos de la financiación, una dimensión prácticamente inexplorada en las Humanidades, han demostrado ser extremadamente valiosos en la valoración de la evolución histórica de una disciplina científica; (3) se han establecido nuevas métricas y valores que constituyen una base de referencia para monitorear la evolución de las HD en España y favorecer las comparaciones, tanto a lo largo del tiempo como
con otros contextos a nivel europeo e internacional.

Referencias bibliográficas

1. Baraibar-Echeverria, Álvaro (2014). In: Baraibar-Echeverria, Álvaro (ed.). Humanidades digitales: Una aproximación transdisciplinar. A Coruña: Universida de da Coruña, SIELAE.
2. González-Blanco, Elena (2013). "Actualidad de las Humanidades Digitales y un ejemplo de ensamblaje poético en la red: ReMetCa". In: Cuadernos hispanoamericanos, n. 761, pp. 53-67. https://doi.org/10.4000/jtei.1274
3. Ortega, Elika; Eunice- Gutiérrez, Silvia (2014). "MapaHD. Una exploración de las humanidades digitales en espanol y portugues". In: Romero-Frías, Esteban (ed.); Sánchez-González. María (ed.). Ciencias sociales y humanidades digitales: técnicas, herramientas y experiencias de e- research e Investigación En Colaboración. La Laguna, Tenerife: Sociedad Latina de Comunicación Social, pp. 101-128.
4. Rojas-Castro, Antonio (2013). "El mapa y el territorio Una aproximación histórico-bibliográfica a la emergencia de las humanidades digitales en España". In: Caracteres, 2, pp. 10-52.

- Romero-Frías, Esteban; Del-Barrio-García, Salvador (2014). "Una visión de lahumanidades digitales a través de sus centros." In: Profesional de la información, v. 23, n. 5, pp. 485– 92. https://doi.org/10.3145/epi.2014.sep.05
- Spence, Paul; González-Blanco, Elena (2014). "A historical perspective on the digital humanities in Spain". In: H- Soz-Kult, November 7th. www.hsozkult.de/debate/id/ diskussionen-2449

Datasets of Criminal Faces Within and Under Facial Recognition Software (FRS) From a Digital Humanities Perspective

Aaron Tucker
artucker@yorku.ca
York University and Ryerson University, Canada

Kieran Ramnarine
kieran.ramnarine@ryerson.ca
Ryerson University, Canada

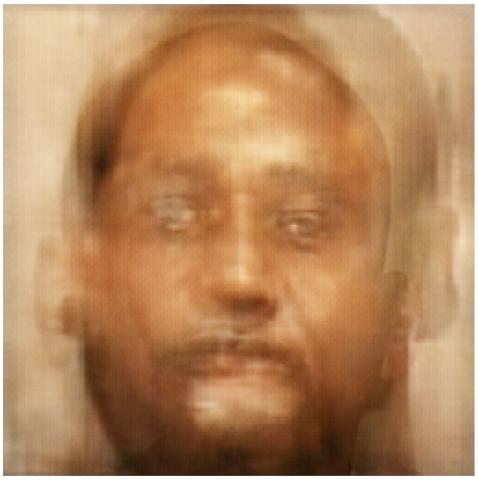
Keywords: machine learning, facial recognition, computer vision, digital literacy, data ethics
Topics: English, North America, 20th Century, Contemporary, artificial intelligence and machine learning, digital activism and advocacy, Communication studies, Media studies

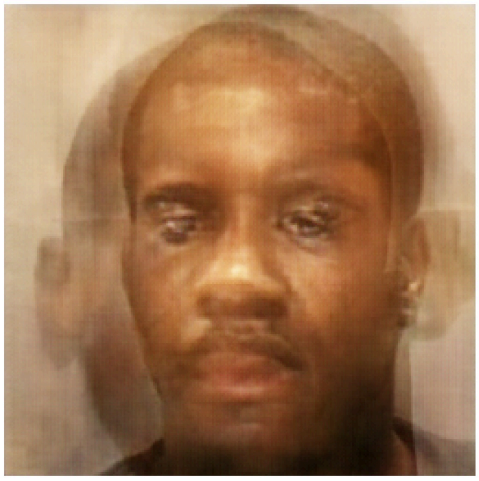
This paper responds to the intersectional problematics of facial databases within contemporary facial recognition software and computer vision machine learning by highlight our ongoing project This Criminal Does Not Exist. Beginning with the MEDS database, our project applies a Convolutional Generative Adversarial Network to produce synthetic faces. Aesthetically, the portraits generated resemble eugencist Francis Galton's "composite portraits" of different races that he deployed in the 19th century. The project is a data visualization project: using machine learning techniques, we have been able to surface what is the "most common" type of face within the dataset; that the portraits generated are primarily of African American males speaks to the types of faces over-represented in these virtual spaces. Further, from this data visualization, "This Criminal Does not Exist" is indicative of contemporary State applications of FRS, bringing to light the clear biases inherent in the dataset, biases further perpetuated through algorithms trained on these types of dataset.

This response is made from a digital humanities perspective that combines principles of ethical data annotation and classification with critical making. In particular, this paper addresses how digital humanities can contribute potential solutions to the ethics of studying and surfacing problematic databases.More specifically, drawing from the the impacts of 19th century pseudo-science like eugenics, phrenology, physiognomy, and signaletics, our project "This Criminal Does Not Exist" signals another potential set of tactics and research creation paths that simultaneously educates the public about the nature of problematic facial datasets, alongside producing arguments about the ethical implications about such databases and their in-built classification practices. Further, this paper explores how digital humanities scholars can provide a public critical engagement with such databases that is grounded in humanized narrative, that does not further replicate and/or ingrain the intersectional and carceral biases of the databases.

Our research begins by recounting how the contemporary study and fears surrounding FRS has been largely focused on large scale corporate- and state-led surveillance apparatuses and their impacts on users' data privacy. This work, exemplified by scholars like Ann Cavoukian and her framework of Privacy by Design, is undeniably useful; similarly, research by surveillance studies theorists like David Lyon and Gary Marx has contributed greatly to advocating for responsible building and application of technologies like FRS. The initial scholarship into the problematic construction of FRS has been driven, in large part, by a wealth of research and reporting about the known inherent biases of the technology, which, as the Georgetown Law Center on Privacy & Technology's report "The Perpetual Line-up" insists, "face recognition may be least accurate for those it is most likely to affect: African Americans." The technology's consistent optimization, in construction and application, for white male faces is especially troubling as the technology moves from being surveilling, national security, and law enforcement tactics, into the ubiquitous, and far more normalized, activities of intervening in job interviews, the monitoring of low incoming housing, and the granting of bank loans. These last three FRS tasks are examples of what Safiya Noble, in her text Algorithms of Oppression, would give as examples of "technological redlining," which she explains is the use of algorithms and big data to "reinforce oppressive social relationships and enact new modes of racial profiling."

Given this, how might digital humanities scholars make the contents of these databases public and available for wider scrutiny and potential regulation while not replicating the dangerous practices that initially led to the construction and implementation of such data? One effective example is artist Trevor Paglen's collaboration with scholar Kate Crawford titled ImageNet Roulette. The project trains an app on the massive ImageNet database's of images labeled in the "person" category. The result is a surfacing of how "ImageNet contains a number of problematic, offensive, and bizarre categories. Hence, the results ImageNet Roulette returns often draw upon those categories. That is by design: we want to shed light on what happens when technical systems are trained using problematic training data." Their accompanying essay, "Excavating AI: The Politics of Images in Machine Learning Training Sets," expands further in labelling their own work as an "archeology of datasets": "we have been digging through the material layers, cataloguing the principles and values by which something was constructed, and analyzing what normative patterns of life were assumed, supported, and reproduced. By excavating the construction of these training sets and their underlying structures, many unquestioned assumptions are revealed." Digital humanities scholars are extremely well suited to take up similar archeological projects, in FRS or other AI- and machine learning-aided environments, as the discipline's focus on ethics, digital tools and humanities-based close-reading techniques grant scholars the abilities to take up the urgent problems of FRS's everyday applications.





Works Cited

Buolamwini, Joy. "Incoding - In the Beginning" Medium. https://medium.com/mit-media-lab/incoding-in-the-beginning-4e2a5c51a45d. May 16 2016. para. 6. Accessed 11 June 2020.

Buolamwini, Joy and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" Conference on Fairness, Accountability, and Transparency. Proceedings of Machine Learning Research 81:1-15, 2018..

Cavoukian, Ann. "The Seven Foundational Principles" https://www.ryerson.ca/pbdce/certification/seven-foundational-principles-of-privacy-by-design/. 2017. Accessed 11 June 2020.

Crawford, Kate and Trevor Paglen. "Excavating AI: The Politics of Images in Machine Learning Training Sets." https://www.excavating.ai/. September 19, 2019. para. 37. Accessed 11 June 2020.

Galton. Francis. "Composite Portraits." Journal of the Anthropological Institute. 1870. 132-144; Hereditary Genius: An Inquiry into its Laws and Consequences. Macmillan and Co, 1869.

Garvie, Clare, Alvaro M. Bedoya and Jonathan Frankle. "The Perpetual Line-Up" Georgetown Law Centre on Privacy and Technology. https//:www.theperpetualineup.org. 2016, para. 23. Accessed 11 June 2020.

Lyon, David. Surveillance as Social Sorting: Privacy, Risk, and Digital Discrimination. Routledge, 2003; Identifying Citizens: ID Cards as Surveillance. Polity, 2009.

Marx, Gary. Windows into the Soul: Surveillance and Society in an Age of High Technology. University of Chicago Press, 2016.

Noble, Safiya U. Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press, 2018.

Potential Archives: How Digital Humanities and Feminist Ethical Praxis Will Transform the Interdisciplinary Artist Archive

Julia G. Polyck-O'Neill
jp03uw@brocku.ca
Brock University, Canada

Keywords: artists' archives, digital archives, multimedia, interdisciplinary art, feminist ethics
Topics: Comparative (2 or more geographical areas), English, North America, 20th Century, Contemporary, digital archiving, public humanities collaborations and methods, Art history, Feminist studies

Potential Archives: How Digital Humanities and Feminist Ethical Praxis Will Transform the Interdisciplinary Artist Archive

As digital media conservators Deena Engel and Glenn Wharton identify in the premise for the Artist Archive Initiative at New York University, conventional approaches to the artist archive neglect to study how the complexity of an artist's interdisciplinary creative practice can confound conventional archival systems and practices. My project demonstrates how artists' archives benefit from non-traditional archival methods that combine emerging digital archival strategies that accommodate and represent community networks and collaborations with the intervention of the artists themselves in the co-creation of accessible multimedia archives. Digital methods will enable artists to augment and customize their archival holdings with attributes such as narratives/narration and networked information.

The need to reconsider material and organizational aspects of artist fonds also has immediate and practical consequences for institutions. Often collections are broken up, and/or the acquisition process can be delayed by technical and policy-driven challenges. National funding opportunities, such as the Canada Council for the Art's Digital Strategy Fund (DSF), add further incentive to the demand for fundamental procedural change.

Archival scholars have identified two interrelated contentions underlying current approaches to artists' archives within the present academic and archival milieu: systemic issues fundamental to archival conventions and practices, and shortcomings of formal organizational strategies within such practices. Feminist archival studies scholars such as Michelle Caswell, Marika Cifor, and Stacy Wood have identified that traditional archival practice is often rooted in colonial and patriarchal cultural and structural conditions. Engel and Wharton, scholars of artists' archives, have addressed how the limitations of conventional archival systems often fail to accommodate the kinds of information, accuracy, and logistical affordances scholars and art professionals require for their research. Specialists in feminist archival studies respond to such organizational shortcomings, observing how the practice of the co-creation of archives with the artist(s) represented within the collections can contribute meaningfully to the value of the collection for scholars and communities. Caswell and Cifor's proposal for a "feminist ethical framework" for archival studies situates the archive socially and culturally, with consideration of relational and affective contexts (24), and Cifor and Wood argue that "critical feminist theory can contribute to existing archival discourse and practice, critiquing concepts that have remained unquestioned, such as community and organization" (3). The addition of autobiographical, narrative, and networked data and digital media forms enable increased access, and have the potential to transform the relationships between artist, archival institution, and user.

This paper explores two main, preliminary ideas: why a transformation of the organization of artist archives is timely and important; and how digital methods and platforms have the potential to benefit artists, arts scholars, and arts archivists. Potential Archives is both a study and a framework, providing both a map of how these non-traditional methods have worked in the past, and a model for how to develop future artist's archives. My study and resulting framework will reconceptualize the interdisciplinary artist archive according to emerging feminist and digital epistemologies and methods to help artists plan for and prepare their future institutional archives and address

emerging needs and concerns, while also assisting arts institutions in addressing such innovations.

Caswell, Michelle. "Teaching to Dismantle White Supremacy in the Archives." The Library Quarterly vol. 87, no. 3, July 2017, pp. 222-235.

Caswell, Michelle and Marika Cifor. "From Human Rights to Feminist Ethics: Radical Empathy in the Archives." Archivaria no. 81, Spring 2016, pp. 23-43.

Cifor, Marika and Stacy Wood. "Critical Feminism in the Archives," in "Critical Archival Studies," eds. Michelle Caswell, Ricardo Punzalan, and T-Kay Sangwand. Special Issue, Journal of Critical Library and Information Studies no. 2, 2017, pp. 1-22.

Engel, Deena and Glenn Wharton. "Managing Contemporary Art Documentation in Museums and Special Collections." Art Documentation: Journal of the Art Libraries Society of North America vol. 36, no. 2, 2017, pp. 293-311.

Analyse automatique pour une étude du genre quels jugements des écrivaines au XIXe siècle ?

Marine Riguet
marineriguet@gmail.com
Labex OBVIL, France

Motasem Alrahabi
Labex OBVIL, France

Keywords: Literary criticism, sentiment analysis, semantic annotation, gender studies, named entity recognition
Topics: Europe, French, 19th Century, semantic analysis, text mining and analysis, Gender and sexuality studies, Literary studies

Cet article présente une méthode d'annotation sémantique développée afin d'étudier le traitement particulier des écrivaines dans la critique littéraire française de la seconde moitié du XIXe siècle. Nous espérons ainsi circonscrire un discours sur la littérature féminine et questionner une pensée littéraire façonnée au prisme du genre. Mais nous entendons également proposer une méthode d'analyse sémantique exportable à d'autres discours, et adaptable aux besoins spécifiques d'autres recherches littéraires.

An Interactive 3D Visualization of RDF-based Digital Editions

Sepideh Alassi
sepideh.alassi@unibas.ch
Digital Humanities Lab, University of Basel

Lukas Rosenthaler
Digital Humanities Lab, University of Basel

Rob Iliffe
Faculty of History, University of Oxford

Keywords: RDF-graph, 3D, visualization, VR, force-directed graph
Topics: Global, English, 15th-17th Century, 18th Century, Contemporary, network analysis and graphs theory and application, semantic analysis, History of science

Humanities research produces a vast amount of data that needs to be visualized to understand and interpret the underlying facts. Modeling data with RDF based OWL ontologies defines a directed graph where nodes are the resources, and the properties the edges. Many digital humanities projects visualize the RDF graphs by flattening them into two dimensions. Although this representation helps researchers with recognizing the direct and indirect connections between the resources, it suffers from loss of information due to the overlap of nodes and edges. One can overcome this problem by visualizing the data as a 3D force graph. An interactive 3D visualization also introduces tangibility to the displayed data so that researchers can rotate the model to study the distribution of the data from every angle. The visualization tool is web-based and connected to a platform that serves the digital editions. This connection enables the users to access the underlying resources by directly clicking on the nodes.

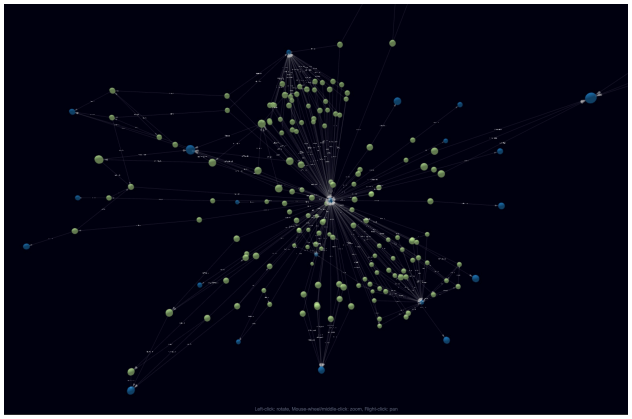


Figure 1: 3D force-directed graph of Newton's scientific correspondence

Representation of the data as a 3D force-directed graph is commonly used for scientific data (Paananen, Wong 2009), but it can be easily adapted to the humanities data as well. To illustrate this, we have chosen to visualize the early modern scientific correspondences. Our database consists of the correspondences of natural philosophers such as Leibniz, Newton, Leonhard Euler, and members of the Bernoulli dynasty. There is already a network that connects the digital editions of these correspondences and makes them openly accessible to the public through one platform (Alassi et al., 2019). This platform is a virtual research environment based on the Knora API (https://www.knora.org/), which manages and stores the data as RDF.Users of this platform can access and query the data online using Angular based front-end components of the API, Knora-ui (https://github.com/dasch-swiss/knora-ui). The 3D visualization tool will be integrated into this user interface and will be openly accessible online. Through Knora, one can derive the graph of the data in the JSON format directly from the triplestore, which is then employed to create a real-time 3D simulation (Figure 1). A configuration step defines graphical features, such as shape, colors, and labels of the nodes and edges (Figure 2). The repulsive forces prescribed on the nodes and edges prevent the overlays of the graph components, and spring-like characteristics of the edges restrain the movement of the nodes. This leads to the formation of clusters of objects that are well connected.

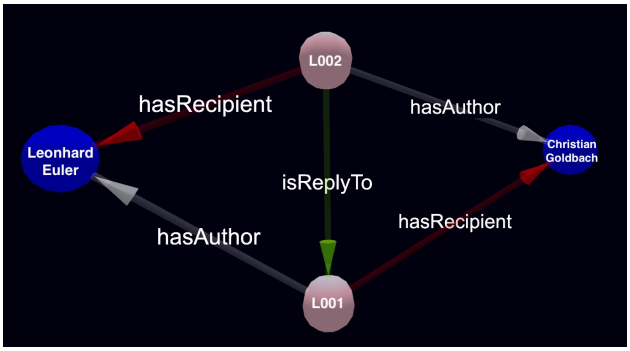


Figure 2: Sample 3D visualization of RDF triples.

Since time plays a crucial role in the study of the historical facts, an additional dimension has to be introduced into the model to represent the time (Schweizer et al., 2015, p.321). In the 3D model, this can be achieved by the dynamic appearance of the nodes relative to the creation date of the letters. All features developed for this tool will be generic and can be used to visualize any RDF based humanities data. We also intend to generate a virtual reality version of the 3D simulations to enhance the interaction of the users with the data.

Bibliography

Paananen, Jussi and Wong, Garry (2009), FORG3D: Force-directed 3D graph editor for visualization of integrated genome scale data, BMC Systems Biology, 2019. DOI:10.1186/1752-0509-3-26

Alassi, Sepideh and Schweizer, Tobias, et al. (2019) Newton virtually meets Euler and Bernoulli, DH2019, https://dev.clariah.nl/files/dh2019/boa/0421.html

Schweizer, Tobias and Rosenthaler, Lukas and Subotic, Ivan (2015) Visualizierung von Annotationen und Verknüpfungen in SALSAH, Geschichte und Informatik, vol.18/19.

Metadata model for an archeological data lake

Pengfei Lui

liu.pengfei@hotmail.fr
Université de Lyon, Lyon 2, ERIC EA 3083, France

Sabine Loudcher

sabine.loudcher@univ-lyon2.fr
Université de Lyon, Lyon 2, ERIC EA 3083, France

Jérôme Darmont
Université de Lyon, Lyon 2, ERIC EA 3083, France

Emmanuelle Perrin
Maison de l'Orient et de la Méditerranée, France

Jean-Pierre Girard
Maison de l'Orient et de la Méditerranée, France

Marie-Odile Rousset
Maison de l'Orient et de la Méditerranée, France

Keywords: archéologie, thésaurus, lac de données, archives ouvertes, métadonnées
Topics: Europe, French, BCE-4th Century, 5th-14th Century, 20th Century, database creation, management, and analysis, systems and information architecture and usability, Archaeology, Computer science

The HyperThesau project was initiated by a multidisciplinary team consisting of two research laboratories of archaeology and computer science, a digital library, two archeological museums and a private company. This project has two main objectives: 1) the design and implementation of an integrated platform to host, search, share and analyze archaeological data; 2) the design of a domain-specific thesaurus taking the whole archaeological data lifecycle into account.

Archeological data may bear many different types (documents, photos, drawings, sensor data, ...). The description of an archaeological object also differs with respect to users, usages and time. Such variety of archeological data induces many scientific challenges related to storing heterogeneous data in a centralized repository, guaranteeing data quality, cleaning and transforming the data to make them interoperable, finding and accessing data efficiently and cross-analyzing the data. To overcome all these challenges, we exploit the concept of data lake.

Haiku Author Recognition

Lubomir Ivanov
livanov@iona.edu
Iona College, United States of America

Keywords: author attribution and identification, haiku, machine learning
Topics: English, North America, 15th-17th Century, 18th Century, 19th Century, artificial intelligence and machine learning, attribution studies and stylometric analysis, Computer science, Literary studies

Haiku Author Recognition

1. Introduction

Haiku is a Japanese poetic form renowned for its brevity and expressiveness. Haiku derives from *renga/renku* - collaborative collections of verses with a 3-line opening *hokku* verse in the form 5-7-5 *on* (equiv. *syllable*). Matsuo Basho made famous the stand-alone *hokku* form, preserving the 5-7-5 on structure. The name *haiku* was associated with this form of *hokku* during 19th century.

Four haiku authors rise in prominence above all: Matsuo Basho (17th century) is considered the "father" of haiku. Yosa Buson (18th century) focused on haiku as an art rather than a reflection of reality. Buson combined hokku with painting, inventing *haiga* (verse-painting). Kobayashi Issa (18-19th century) reinvented haiku through his depth of feeling and humanism. In the second half of the 19th century, Masaoka Shiki critically re-evaluated the art of haiku (coining the term), braking away from the traditional 5-7-5 form, and popularizing the poetic style beyond Japan.

We present a study, which employs authorship attribution techniques to determine the distinctiveness of poetic styles in haiku, focusing on the poetry of Basho, Buson, Issa, and Shiki. There has been little work in the field of haiku attribution. A theoretical study of phonological complexity in haiku was presented in [1]. An approach to automatic evaluation of the quality of haiku was presented in [2]. An interesting work [3] deals with identifying unintended haiku in text. We approach haiku attribution as a classification problem: Given a set of attributed haikus, we train classifiers to recognize the writing style of each poet, and apply an ensemble of trained models to unattributed texts.

2. Our Haiku Corpus

The first step in creating our model was obtaining a haiku corpus. There are three approaches:

- Use actual haikus written in hiragana (a form of Japanese alphabet)
- Use Roman alphabet transcriptions (rōmaji) of haikus.
- Use English translations of haikus.

While using hiragana haikus is arguably the best option, our software lacks the capability to process hiragana text. English translations of haikus are readily available, but while research suggests that the authorial signal is stronger than the translators' [4], we do not know if that applies to haiku. We opted to construct a corpus of romaji transcribed haikus. This was difficult

since most resources are either hiragana originals or translations. We obtained 723 haikus by Basho from [5], 842 haikus by Buson from [6], and 603 haikus by Issa from [7]. Finding transcribed Shiki haikus proved extremely challenging. Even though Shiki wrote over 24000 haikus, only a handful have been transcribed into romaji. Failing to secure transcriptions, we downloaded the full set of 24000 hiragana haikus from [8]. We then used an online hiragana-to-rōmaji transcription tool [9] to transcribe 967 randomly selected haikus by Shiki. Since many of the extracted haikus were organized alphabetically or by topic, we wrote Python code to randomly shuffle the order of the haikus for each author. A different program broke up the haikus into files of size 50 haikus each.

3. Attribution Methodology

Our attribution software is a based on JGAAP [10] and implements an ensemble of classifier/stylistic-feature pairs [11,12]. For this study, we limited the set of stylistic features to character-2/3/4/5-grams (CnG), word-2/3-grams (WnG), vowel-initiated words (VIW), and first-word-in-sentence (FWIS). The classifiers used were support vector machines with sequential minimal optimization (SMO) and multilayer perceptrons (MLP).

4. Results

We conducted several experiments, where we randomly chose one 50-haiku file for each author and removed it from the training set. We trained the classifiers on the remaining set of haikus using leave-one-out (L1O) validation. The results of the training for three sets of experiments are presented in Table 1:

	Experiment 1	Experiment 2	Experiment 3
C2G/MLP	91.80%	93.44%	93.44%
C3G/MLP	95.08%	96.72%	93.44%
C4G/MLP	90.16%	100%	95.08%
C5G/MLP	93.44%	95.08%	93.44%
FWIS/MLP	83.61%	80.33%	88.52%
VIW/MLP	93.44%	93.44%	95.08%
W2G/MLP	75.40%	72.13%	78.69%
W3G/MLP	80.33%	80.33%	77.05%
C2G/SMO	93.44%	91.80%	93.44%
C3G/ SMO	96.72%	96.72%	96.72%
C4G/ SMO	88.52%	100%	95.08%
C5G/ SMO	93.44%	95.08%	93.44%
FWIS/ SMO	77.05%	77.05%	83.61%
VIW/ SMO	95.08%	93.44%	96.72%
W2G/ SMO	68.83%	77.05%	72.13%
W3G/ SMO	80.33%	83.61%	80.33%
Ensemble	100.00%	100.00%	100.00%

Table 1: Training Accuracy for Basho, Buson, Issa, and Shiki

Next, we tested the authorship of the 50-haiku files that were left out of the training. The results of those experiments are presented in Table 2:

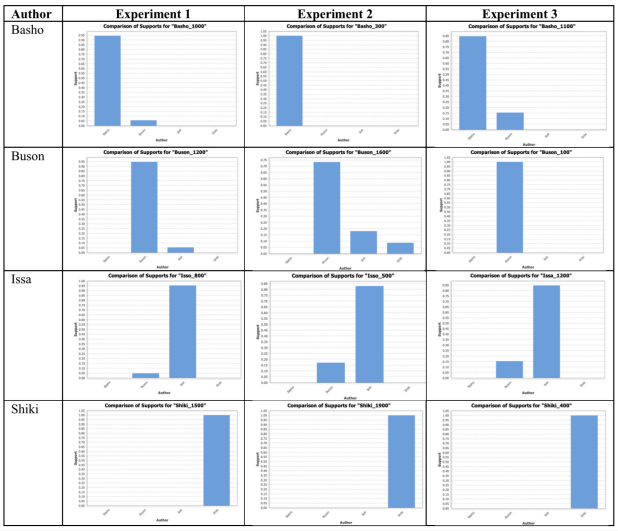


Table 2: Attribution Results for Basho, Buson, Issa, and Shiki

It is quite clear that even with a reduced set of stylistic features, the attribution is very strong and the author identification definitive. We conducted an additional set of experiments, where we used each of the trained models to test the authorship of five haikus by the 18th century haiku poet Takarai Kikaku. The models were not trained on Kikaku, so, as expected, the results were split among two or more authors (Table 3):

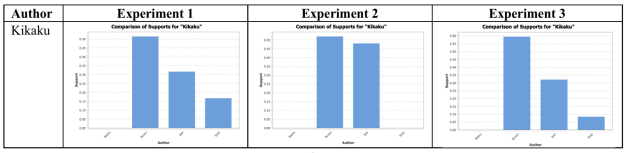


Table 3: Attribution Results for Kikaku

Interestingly, Kikaku was a prominent student and disciple of Basho, yet none of the models makes that association. This is most likely due to the small number of Kikaku haikus tested.

5. Conclusion and Future Work

We presented results from haiku author identification experiments, which suggest that haiku authorship can be determined even with a limited set of stylistic features from rōmaji-transcribed haikus. Our next efforts will be to experiment with a larger set of haiku authors, with English translations, and, possibly, with hiragana haikus. Among the questions we wish to answer are:

- What is the minimal set of haikus sufficient to identify an author?
- Is the authorial signal stronger than the translator's for haiku translations?
- Can prosodic features be used for haiku author identification?
- Does the historical period affect the accuracy of attribution?

References:

[1] Hayata, K. (2018). "Phonological Complexity in the Japanese Short Poetry: Coexistence Between Nearest-Neighbor Correlations and Far-Reaching Anticorrelations". Front. Phys. 6:31. doi: 10.3389/fphy.2018.00031

[2] Kikuchi, S. et al. (2016). "Quality estimation for Japanese Haiku poems using Neural Network". IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, DOI: 10.1109/SSCI.2016.7850030

[3] Online resource: https://labs.ft.com/2016/07/finding-hidden-haiku/

[4] Hoover, D. (2019). "The Invisible Translator Revisited". Digital Humanities Conference (DH2019). Utrecht, The Netherlands

[5] Barnhill, D.L. (2004). "Basho's Haiku: Selected Poems of Matsuo Basho". State University of New York Press, ISBN-13: 978-0791461662

[6] Persinger, A. (2013). "Foxfire: The Selected Poems of Yosa Buson, a Translation". Theses and Dissertations. Paper 748.

[7] Terebes, G. (2000). "A cup of tea with Isa". Orpheusz Publishing House, Budapest, 2000, c. bilingual, corrected and expanded version of volume 2.

[8] Online resource: https://terebess.hu/english/haiku/shiki.html

[9] Online resource: http://www.romajidesu.com/translator

[10] Juola, P. 2009. JGAAP: A system for comparative evaluation of authorship attribution. Journal of the Chicago Colloquium on Digital Humanities and Computer Science, 1(1): 1-5.

[11] Petrovic, S., Berton, G., Campbell, S., Ivanov, L. 2015. Attribution of 18th Century Political Writings Using Machine Learning. Journal of Technologies in Society, volume 11, issue 3, pp. 1-13

[12] Petrovic, S., Berton, G., Schiaffino, R., Ivanov, L. 2016. Examining the Thomas Paine Corpus: Automate d Computer Author Attribution Methodology Applied to Thomas Paine's Writings. Chapter, New Directions in Thomas Paine Studies, Edition: 1, Publisher: Palgrave Macmillan US, Editors: S. Cleary I. Stabell, DOI: 10.1057/9781137589996

"Open GLAM": Opening up digital cultural heritage collections for the digital humanities

Roger Christopher Gillis
roger.gillis@dal.ca
Dalhousie University

Keywords: Copyright, Cultural Heritage, Digital Collections, Open Access, Open GLAM
Topics: Global, English, Contemporary, copyright, licensing, and permissions standards, systems, and processes, digital access, privacy, and ethics analysis, Galleries and museum studies, Library & information science

Over the past twenty plus years, cultural heritage organizations have put vast amounts of digitized available online.The movement towards Open Access for digitized cultural heritage has come to be known as Open GLAM (Galleries, Libraries, Archives, and Museums). This presentation will explore key issues and recent efforts around Open GLAM and the broader issues of Open Access to cultural heritage, and in particular will highlight many current activities around Open GLAM, as well as issues and challenges that exist in this area.

Towards Modeling the European Novel. Introducing ELTeC for Multilingual and Pluricultural Distant Reading

J. Berenike Herrmann
berenike.herrmann@unibas.ch
University of Basel, Switzerland

Carolin Odebrecht
Humboldt-University Berlin, Germany

Diana Santos
University of Oslo, Norway

Pieter Francois
University of Oxford, UK

Keywords: corpus, TEI, literary modeling, comparative literary studies, 19th Century
Topics: Comparative (2 or more geographical areas), Europe, English, 19th Century, data modeling, database creation, management, and analysis, Humanities computing, Literary studies

Towards Modeling the European Novel. Introducing ELTeC for Multilingual and Pluricultural Distant Reading

This contribution reports on the collaborative effort of building an open access multilingual corpus of European novels published 1840-1920 (the European Literary Text Collection - ELTeC) within the COST Actio n "Distant Reading".[1]

Working at the intersection of many languages and cultures, we address practical and technical aspects of corpus design based on a theoretical discussion of pluri-cultural computational modeling of literature. In the corpus design, we adopt a metadata-based approach that allows for representing the diversity of novels published 1840-1920 across Europe. Our sampling and balancing criteria use metadata including publication date, text length, reprint counts and authors' gender,[2] and we deliberately focus on inclusion of non-canonical novels.

We have built a workflow for systematically sampling and encoding novels as well as a consistent annotation model of data and metadata (cf. Burnard, Schöch, Odebrecht, 2019). Currently, ELTeC constitutes a dynamic intersection of fictional discourse in fourteen languages, including Czech, English, French, German, Greek, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovenian, and Spanish (ca. 600 text candidates amounting to ca. 52 mio. words).[3] Figure 1 depicts the current state of the Portuguese sample, across categories "text length" (long, medium, short), "gender" (female, male), and "date slot" (T1-4; https://distant reading.github.io/ELTeC/por/index.html).

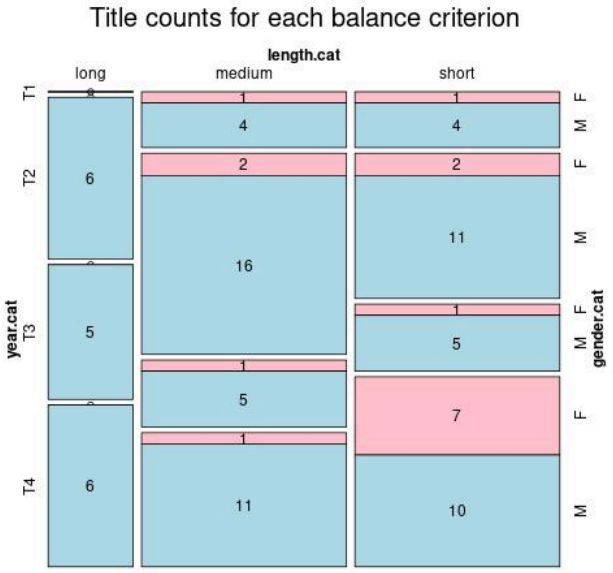


Figure 1. Screenshot of Portuguese sample

Our TEI-XML (TEI Consortium, 2019) encoding scheme aims is minimal, but aims at facilitating a rich and well-informed distant reading. ELTeC is rooted in the open data movement, with collaborative data creation and an open access extensive documentation for (meta-)data schema, decisions and workflows.[4] Each version of ELTeC is archived via Zenodo.[5] Thus, our (meta)data are re-usable, interoperable, accessible and findable (cf. FAIR Guiding Principles; Wilkinson et al., 2016).

In view of the problematic notion of "representativeness" (see Biber, 1993), ELTeC deliberately refrains from modeling the statistical distribution of populations of publication or reception (cf. Herrmann & Lauer, 2019). Rather, we address the inevitable bias included in the sampling (see Bode, 2018), as well as the explicit link to research questions (Underwood, 2019; Lüdeling, 2011) and the act of construction (Piper, 2019).

ELTeC caters to the computational modeling of literature at the intersection of cultures, nations, languages, genders, but also poetologies, trends, and traditions, in a historical period of extreme aesthetic change and diversity. Giving one example, in collaboration with other working groups, using demonym and named entity recognition, it is possible to comparatively explore images of 'the other' (ethnic, national, regional; Leerssen, 2016). Generally, in the creation of ELTeC we aim at inductively defining what a 'novel' is, allowing for diverse approaches in literary theory and history to be explored and tested.

The research described in this paper was conducted in the context of the COST Action "Distant Reading for European Literary History" (CA16204 - "Distant-Reading"). Find out more at: http://www.distant-reading.net. COST is funded by the Horizon 2020 Framework Programme of the EU.

References

Biber, D. (1993). Representativeness in Corpus Design. Literary and Linguistic Computing (8), 243-257.

Bode, K. (2018). A World of Fiction: Digital Collections and the Future of Literary History. University of Michigan Press.

Burnard, L., Schöch, C., & Odebrecht, C. (2019). In search of comity: TEI for distant reading. Book of Abstracts TEI Conference Graz 2019. http://gams.uni-graz.at/context:tei2019

Herrmann, J. B., & Lauer, G. (2018). Korpusliteraturwissenschaft. Zur Konzeption und Praxis am Beispiel eines Korpus zur literarischen Moderne. Osnabrücker Beiträge Zur Sprachtheorie (OBST), 92, 127-156.

Leerssen, J. (2016). Imagology: On using ethnicity to make sense of the world. Iberic@l, Revue d'études ibériques et ibéro-américaines, 10, 13-31. http://iberical.paris-sorbonne.fr/wp-content/uploads/2017/02/Pages-from-Iberic@l-no10-automne-2016-Final-2.pdf

Lüdeling, A. (2011). Corpora in Linguistics: Sampling and Annotation. In K. Grandin (ed.), Going Digital. Nobel Symposium, 147, (220-243). New York: Science History Publications.

Piper, A. (2017). Think Small: On Literary Modeling. PMLA, 132(3), 651-658. https://doi.org/10.1632/pmla.2017.132.3.651

TEI Consortium (eds). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.6.0. [2019]. http://www.tei-c.org/Guidelines/P5/ (2019-09-25).

Underwood, T. (2019). Distant horizons: Digital evidence and literary change. Chicago: University of Chicago Press.

[1] https://www.distant-reading.net/

[2] See our white paper on https://distantreading.github.io/sampling_proposal.pdf

[3] Current state of ELTeC: https://distantreading.github.io/ELTeC/index.html

[4] Corpus data: https://github.com/COST-ELTeC, documentation: https://github.com/distantreading/WG1

[5] https://zenodo.org/communities/eltec

AI4AV (Artificial Intelligence for Audiovisual): Design and Evaluation of a Shared System for LAMs

Maria Esteva
Texas Advanced Computing Center

Tanya Clement
tclement@utexas.edu
Department of English

Weijia Xu
Texas Advanced Computing Center

Choate Aaron
UT Libraries

Hannah Robbins Hopkins
Department of English

Keywords: audiovisual, artificial intelligence, machine learning, professional values, shared infrastructure
Topics: South America, English, North America, Contemporary, artificial intelligence and machine learning, metadata standards, systems, and methods, Computer science, Library & information science

Audiovisual (AV) materials are predominant historical and scientific records of our times, and their numbers are increasing exponentially in collecting institutions. Tasked with preserving and making AV materials available, libraries, archives, and museums (LAMs), need to find efficient and scalable curation solutions. Using machine learning (ML) to generate metadata is promising, but to adopt such methods information professionals must overcome a host of technological and cultural challenges. We introduce the AI4AV project in which we are conducting research around the design and evaluation of a system (currently a prototype) that uses ML to translate audio to text as well as natural language processing to classify and describe AV materials within open computing infrastructure that can be shared by multiple LAMs. This

presentation describes the testbed collection, the ML and NLP methods and computing resources, and the protocol to incorporate LAMs values in the design and evaluation of the system.

Developing the Data Foundry: the National Library of Scotland's data-delivery platform

Sarah Ames
sarah.ames@nls.uk
National Library of Scotland

Keywords: libraries, data, open data, rights
Topics: Global, Europe, English, Contemporary, data publishing projects, systems, and methods, digital libraries creation, management, and analysis, Book and print history, Library & information science

The Collections as Data movement has gained significant traction in recent years, with large-scale projects leading the way in shaping and advocating for best practice (Padilla et al 2019). These studies, along with the OpenGLAM movement, have encouraged cultural heritage organisations to make collections available in machine readable formats and to support computational research with the collections, enabling libraries to cast new light on collections and present them in new ways for digital humanities audiences.

However, while there have been a number of recent, essential studies around Collections as Data, as well as research into making collections available openly and the reasoning behind this (Pekel 2014, Terras 2015), there has been little to date from an institutional point of view about what is involved in opening up the collections in this way.

How can libraries open up collections to wider audiences? How do we turn collections into data? What challenges does this present, relating to rights, access, and data management? What ethical considerations are needed and how can libraries be transparent about decision-making processes as they generate increasing amounts of data, becoming 'producers' of their own collections?

This paper lifts the lid on the process of making data available in a national library context and considers the changes to existing activities, processes and outlook in releasing collections as data.

The National Library of Scotland launched the Data Foundry (https://data.nls.uk/) in September 2019. As part of the Library's Digital Scholarship Service, the Data Foundry provides access to data collections including digitised collections; metadata; map and spatial data; and organisational data; with further collections such as web archive and audiovisual data planned for future release.

The Data Foundry is based on three core principles: open, transparent and practical (National Library of Scotland 2019 [1]). The platform was designed to be a clear, easy-to-use website, with tiered data downloads; clear rights information; and at-a-glance details contextualising the datasets.

Collections on the Data Foundry are published openly, in reusable formats, and the Library does not assert further copyright over the datasets it produces (National Library of Scotland 2019 [2]). Furthermore, with transparency a key principle, the Data Foundry provides information about data provenance and the reasons behind why and how certain items have been digitised and 'turned into' data above others (Ames 2019).

Producing the Data Foundry has been a Library-wide effort. Working at the intersection of collections, technology and research, the National Library of Scotland's Digital Scholarship Service draws upon existing expertise across the Library - including Rights, Developers, Curators, Metadata - as well as working closely with researchers to understand their needs.

This paper will highlight the practical side of opening up library collections for digital humanities use, exploring the everyday challenges and obstacles such as rights and technical issues and changes to workflows required to produce collections as data, as well as the broader implications of making collections available at scale for libraries and their users.

References

Ames, Sarah. (2019). 'Digital scholarship and data provenance at The National Library of Scotland'. Libraries as Research Partner in Digital Humanities. http://doi.org/10.5281/zenodo.3269291

National Library of Scotland [1] (2019). https://data.nls.uk/

National Library of Scotland [2] (2019). National Library of Scotland Open Data Publication Plan. https://data.nls.uk/download/national-library-of-scotland-open-data-publication-plan.pdf

OpenGLAM. (2019). 'Ways forward to Open Access for cultural heritage'. OpenGLAM. https://openglam.org/2019/04/30/openglam-principles-ways-forward-to-open-access-for-cultural-heritage/

Padilla, Thomas, Allen, Laurie, Frost, Hannah, Potvin, Sarah, Russey Roke, Elizabeth, & Varner, Stewart. (2019). Final Report --- Always Already Computational: Collections as Data (Version 1). http://doi.org/10.5281/zenodo.3152935

Padilla, Thomas, Scates Ketler, Hannah, Allen, Laurie, Varner, Stewart. (2019). Collections as Data: Part to Whole. https://collectionsasdata.github.io/part2whole/

Pekel, Joris. (2014). Democratising the Rijksmuseum. Europeana Foundation. https://pro.europeana.eu/files/Europeana_Professional/Publications/Democratising%20the%20Rijksmuseum.pdf

Terras, Melissa. (2015). 'Opening Access to Collections: the Making and Using of Open Digitised Cultural Content'. Online Information Review. http://discovery.ucl.ac.uk/1469561/1/MelissaTerras_OpeningAccess_OIR.pdf

The "No-Interface" Interface for Research VR

Andrew Burrell
University of Technology, Sydney

Rachel Hendery
r.hendery@westernsydney.edu.au
Western Sydney University, Australia

Keywords: Virtual reality, interfaces, design, linguistics, computer history
Topics: Global, English, 20th Century, Contemporary, Interface design, development, and analysis, virtual and augmented reality creation, systems, and analysis, Design studies, Linguistics

The potential to create innovative and immersive systems for interaction within Virtual reality (VR), as suggested by VR pioneers, initially appeared limitless, and promised to bring rise to a whole new paradigm of interface and interaction design. Yet over quarter of a century later, in the so-called second-wave of VR, most of the standard metaphors for creating interfaces in VR are an extension of either the desktop metaphor of the personal computer or the Heads Up Display of the first person computer game. In this paper we consider how we then design for interactions in VR that take full advantage of the media's affordances, without relying on these familiar metaphors, yet being familiar enough as to engender intuitive and meaningful use? We will explore this question by introducing our Layered Horizons information visualisation as a case study alongside a survey of important voices from the history of VR interface design.

Mapping Meaning: learnings from indigenous mapping technology for Australia's digital humanities mapping infrastructure

Bill Pascoe
bill.pascoe@newcastle.edu.au
University of Newcastle, Australia

Keywords: digital mapping, indigenous, ontology, ethics, meaning
Topics: Global, English, Australia/Oceania, BCE-4th Century, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), spatial & spatio-temporal analysis, modeling and visualization, First nations and indigenous studies, Geography and geo-humanities

Time Layered Cultural Map (TLCMap) is an ambitious, ARC funded, digital humanities mapping infrastructure initiative in Australia. It aims to make open access digital mapping easier and develop new functionality for humanities researchers, supporting open scholarship. TLCMap is at the intersection of digital humanities research, the general public, colonial histories and 40,000 years of living indigenous culture in which 'country' is of central importance. To avoid symbolic repetition of the violence of 'terra nullius' this Australian mapping infrastructure begins with learning from rather than about indigenous mapping technology. Recent indigenous mapping projects highlight the usefulness of digital mapping, and areas for improvement in functionality. Beyond mapping indigenous 'content' as object of study, our development is informed by indigenous mapping techniques. Theoretical considerations about the ontological and ethical functions of mapping and translation lead to practical digital mapping tools.

Unsupervised Rumor Detection on Twitter using Topic Homogeneity

David Lawrence Shepard
shepard.david@gmail.com
UCLA, United States of America

Takako Hashimoto
Chiba University of Commerce, Japan

Kilho Shin
University of Hyogo, Japan

Takeaki Uno
National Institute of Informatics, Japan

Tetsuji Kuboyama
Gakushuin University, Japan

Keywords: twitter, text mining, graph theory, rumor detection
Topics: Asia, Global, English, Contemporary, social media analysis and methods, text mining and analysis, History, Humanities computing

Rumor detection on twitter is a highly-studied topic. Our prior work showed that it is possible to identify rumors (stories about events that have no basis in fact) on twitter by observing the relationship between the number of tweets about an event and the number of topics within those tweets (Shepard et al., 2019). In that paper, we presented case studies, but could not measure the difference between rumors and facts. This paper builds on that work by presenting a formula for making that distinction: given a data set of tweet texts with time stamps, our method automatically discovers rumors in those tweets. This paper presents our method, which is completely unsupervised in contrast to most recent work on rumor detection. As a case study, it demonstrates the method's effectiveness on a sample dataset of tweets from the 3/11 earthquake in Japan.

Background

At least seven new methods have been proposed for rumor detection on twitter in the past two years, including (Wang et al., 2017; Singh et al., 2017; Chen et al., 2017; Yoshida and Aritsugi, 2019; Ma et al., 2018; Poddar et al., 2018; Lin et al., 2019). One major constraint of these approaches, however, is that all of them require supervised learning. Our major intervention is that our approach is unsupervised and language-independent. Our algorithm makes only the assumption that the text has been tokenized into words.

Method

Our approach results from the observation that little new information emerges about rumors. For example, if a chemical spill is announced on twitter, people often retweet the announcement. People are also likely to be affected in some way, such as seeing emergency vehicles or evacuating the area. New information will be created as the event progresses. If the spill is just a rumor, however, people will have no new information to add, and so will simply retweet the original rumor. We call the amount of information surrounding an event its "topic diversity." Calculating an event's topic diversity allows us to classify an event as a rumor or a fact. A true event will have both high tweet counts and high topic diversity, while a rumor will have high tweet counts but low topic diversity.

The first step in calculating topic diversity is separating our data into what we call word-window tweet lists, or WWTLs. These are lists of tweets that mention a word, split into 30-minute windows. We generate WWTLs for every word in our dataset of tweets, after removing stopwords. Next, we compute each WWTL's topic diversity by performing topic modeling on each WWTL and counting the number of topics produced. We use a novel method of topic modeling that determines the number of topics automatically based on maximal clique enumeration (MCE). To prepare a TTWL for MCE, we first generate a graph of all tweets. Each node represents one tweet, and edges are generated by computing the Jaccard similarity of each pair of tweets based on shared vocabulary. Then, we perform data polishing (Uno et al. , 2015; Uno et al., 2017) and MCE using nysol_python (nysol, 2019; Nakamoto and Hamuro, 2018) to produce a clique count for each WWTL. A WWTL with a large number of cliques has high topic diversity.

Finally, we examine whether increases in tweet counts in WWTLs correlate with topic diversity. For each word, we find the WWTL with the maximum tweet count (Tmax). Then, we find the minimum number of non-zero tweets prior to that maximum (Tmin), and call that period between those windows a "keyword rise." For each keyword rise, we find the number of communities at the beginning and the end of the keyword rise (CTmin and CTmax). We then compute what we call the topic homogeneity factor (THF), using the following formula:

THF = ((T_max - T_min) / (T_min)) / ((C_T_max - C_T_min) / (C_T_min))

We find that keywords with a THF of 25 or greater are likely to be rumors.

Experiment

To validate our method, we performed an experiment on a data set of around 200 million tweets sent in the three weeks after the 3/11 earthquake in Japan, gathered by the social media company Hottolink (https://www.hottolink.co.jp/english). We knew a number of rumors had spread during the disaster, but were later corrected in official news sources.

Our test focused on identifying keywords related to two rumors. The first was about a chemical leak at the Cosmo Oil plant. The second was a widely-retweeted message purporting to be the last words of a dying system administrator, referred to as the "Geek House Asakusa" tweet.

To compare the patterns of these known rumor keywords against words not associated with rumors, we arbitrarily selected 150 other words that we expected to exhibit a diverse set of frequencies and topic diversity. We included users' twitter handles, government agencies, and locations in the affected area. We expected that twitter handles would be mentioned infrequently, and have low topic diversity, while government agencies and affected areas would be mentioned frequently and have high topic diversity. Given the variety among these words' frequency and topic diversity, we expected that many of them would have either word frequency or topic diversity measures that could confuse our method.

Results

Using our method, it was easy to differentiate words related to rumors from other words. The term "コスモ" ("Cosmo") had a THF of 39, and "コスモ石油"1 ("Cosmo Oil") had a THF of 73. Similar results were obtained for "ギーク" ("geek"), with a THF of 41.73. Of the 150 non-rumor keywords, we found that with a threshold of 25.0, only one word was incorrectly classified as a rumor. Given this success, we considered our method to be effective.

Conclusion

We have shown that our method can effectively detect keywords likely to be associated with rumors. While we acknowledge that it was tested on a dataset from the past and not on a live tweet stream, in future work, we plan to experiment on a live tweet stream. We also plan to experiment with political rumors likely to generate discussion even if they are not true, which would produce higher topic diversity. Overall, though, we found our method to be effective at detecting rumors in our dataset and anticipate it to be effective at discovering rumors in similar dataset.

Works Cited

Chen, Y.-C., Liu, Z.-Y. and Kao, H.-Y. (2017). IKM at SemEval-2017 Task 8: Convolutional neural networks for stance detection and rumor verification. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 465-469.

Lin, X., Liao, X., Xu, T., Pian, W. and Wong, K.-F. (2019). Rumor Detection with Hierarchical Recurrent Convolutional Neural Network. In Tang, J., Kan, M.-Y., Zhao, D., Li, S. and Zan, H. (eds), Natural Language Processing and Chinese Computing. (Lecture Notes in Computer Science). Springer International Publishing, pp. 338-48.

Ma, J., Gao, W. and Wong, K.-F. (2018). Rumor detection on twitter with tree-structured recursive neural networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1980-1989.

Nakamoto, M. and Hamuro, Y. (2018). NYSOL: Preprocessing Tool for Big Data in Python. Fukuoka, Japan https://www.ieice.org/publications/conference-FIT-DVDs/FIT2018/data/pdf/B-015.pdf (accessed 12 October 2019).

Nysol (2019). Nysol/Nysol_python. C https://github.com/nysol/nysol_python (accessed 12 October 2019).

Poddar, L., Hsu, W., Lee, M. L. and Subramaniyam, S. (2018). Predicting Stances in Twitter Conversations for Detecting Veracity of Rumors: A Neural Approach. 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, pp. 65-72.

Shepard, D., Hashimoto, T., Shin, K., Uno, T. and Kuboyama, T. (2019). The Narrow Scopes of Fake News: Detecting Fake News Using Topic Diversity Measures. Utrecht (accessed 12 October 2019).

Singh, V., Narayan, S., Akhtar, M. S., Ekbal, A. and Bhattacharyya, P. (2017). IITP at SemEval-2017 Task 8: A supervised approach for rumour evaluation. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 497-501.

Uno, T., Maegawa, H., Nakahara, T., Hamuro, Y., Yoshinaka, R. and Tatsuta, M. (2015). Micro-Clustering: Finding Small Clusters in Large Diversity. ArXiv:1507.03067 [Cs] http://arxiv.org/abs/1507.03067 (accessed 13 October 2019).

Uno, T., Maegawa, H., Nakahara, T., Hamuro, Y., Yoshinaka, R. and Tatsuta, M. (2017). Micro-clustering by data polishing. 2017 IEEE International Conference on Big Data (Big Data). pp. 1012-18 doi:10.1109/BigData.2017.8258024.

Wang, F., Lan, M. and Wu, Y. (2017). ECNU at SemEval-2017 Task 8: Rumour evaluation using effective features and supervised ensemble models. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 491-496.

Yoshida, Z. and Aritsugi, M. (2019). Rumor Detection in Twitter with Social Graph Structures.

Third International Congress on Information and Communication Technology. Springer, pp. 589-598.

From Silo to Repo: Enforcing File Structure to Improve Workflow and Access

Jessica Dussault
jdussault@unl.edu
University of Nebraska-Lincoln, United States of America

Laura K Weakly
lweakly2@unl.edu
University of Nebraska-Lincoln, United States of America

Keywords: project management, repositories, workflow, data management
Topics: English, North America, Contemporary, data publishing projects, systems, and methods, project design, organization, management, Humanities computing, Library & information science

From Silo to Repo: Enforcing File Structure to Improve Workflow and Access

Dussault, Jessica; Weakly, Laura K.University of Nebraska-Lincoln

The University of Nebraska-Lincoln's Center for Digital Research in the Humanities (CDRH) has created more than 80 projects since 1998.[1] Because of the collaborative and experimental nature of Digital Humanities projects, a variety of technologies and structures were utilized, which has proved challenging to support as we continue to develop projects.[2] Recently the CDRH has implemented policies that separate data from websites and standardize file locations. These policies have been reinforced by building software with expectations about data location, which makes creating, maintaining, and upgrading sites more manageable. This paper examines the structural changes of organizing files, the impact on workflow, and the increased accessibility of data for researchers.

File Organization

The CDRH's largest and most well-known projects are document-driven digital archives, producing tens of thousands of encoded TEI-XML[3] files. Historically, these files were part of a website's code base, but projects shared no common structure. The first step towards managing data for dozens of projects was to move the documents out of website code into "data repositories."[4] Datura[5], CDRH-developed software, provides customizable scripts for transforming a data repository's documents into other formats, such as HTML or JSON for search indexes. The location of documents is enforced by Datura's expectations of file types.

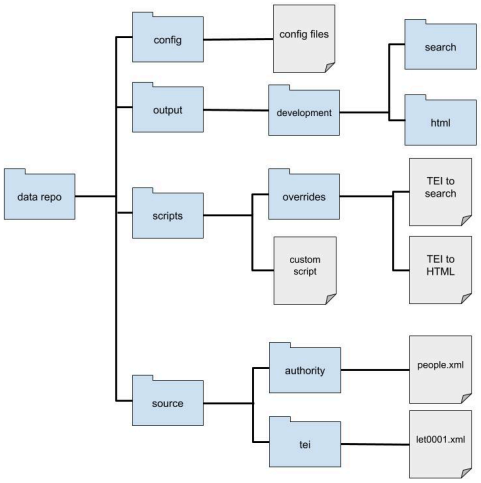


Figure 1: The basic structure of a data repository. The source documents, stored in subdirectories by type, are transformed by scripts which apply overrides and configuration settings into files in the output directory. Most commonly, TEI files are transformed into HTML for display or JSON for populating search indexes.

Images are stored separately from the website served with an International Image Interoperability Framework-compatible image server.[6] These changes to data and image locations allow us to focus on standardizing other aspects of our websites, such as publishing documents with a center-wide API.

Workflow

This Datura-driven standardization has led to improvements in project workflow. Students and other researchers can be commonly trained on file naming and data structures. They are trained to use a version control system, currently git hosted on GitHub, allowing easier access to file sharing and virtually eliminating data being lost or overwritten. Previously students working on multiple projects needed training on each project as to where metadata files and images were stored, scan specifications, file naming, and file sharing. Now students can work across projects with little to no learning curve for each new project.

Increased Accessibility

The CDRH has a commitment to making data available for scholars. We have long exposed XML source files powering individual pages[7], but the new data repository system makes downloading and working with project files even easier. The CDRH publishes data repositories on the code-sharing platform GitHub, tagged with "tei-xml" to aid discoverability.[8] These public repositories can be downloaded or referenced remotely. An artist has already used repository documents to power a small site.[9] The public repositories also provide student contributors with a resource to showcase their project work.[10]

Conclusion

Though this system has drawbacks that we will discuss during our presentation, our data-first approach has dramatically decreased development time spent on creating websites and on student training. Gathering documents together has improved the internal workflow of collaborators, laid the groundwork for future initiatives to renovate aging websites, and aided in the dissemination of data and scholarship to researchers around the globe.

Footnotes

1 https://cdrh.unl.edu/about/creatingcdrh#establishment

2 Discussion around designing Digital Humanities projects to enable future support is increasingly important in the discipline. See http://www.digitalhumanities.org/dhq/ vol/13/1/000411/000411.html

3 https://tei-c.org/

4 Typically, each project has a single data repository, but in the case of larger and more complex projects like The Walt Whitman Archive, a project may have multiple data repositories to group documents with different concerns like marginalia or scribal documents.

5 https://github.com/CDRH/datura

6 We use the Cantaloupe server (https://cantaloupe-project.github.io/) with International Image Interoperability Framework Image API (https://iiif.io/api/image/2.1/)

7 The Willa Cather Archive is an early example of this practice and continues to provide links to source files on individual document pages: https://cather.unl.edu/

8 https://github.com/search?q=topic%3Atei-xml+org%3ACDRH&type=Repositories

9 https://love-and-seduction.glitch.me/ by John Skiles Skinner

10 For more on the complexities of properly crediting student work, see http://www.digitalhumanities.org/dhq/vol/11/3/000322/000322.html

Works Cited

Cantaloupe Project (2019). "Cantaloupe." University of Illinois. https://cantaloupe-project.github.io/

Center for Digital Research in the Humanities (2019). "CDRH Organization." GitHub. https://github.com/cdrh/

Center for Digital Research in the Humanities (2019). "Creating the CDRH." University of Nebraska-Lincoln. https://cdrh.unl.edu/about/creatingcdrh#establishment

Center for Digital Research in the Humanities (2019). "Datura." GitHub. https://github.com/CDRH/datura

Dalziel, Karin; Dussault, Jessica; Tunink, Greg (2018). "Legacy No Longer: Designing Sustainable Systems for Website Development." DH2018. https://dh2018.adho.org/en/legacy-no-longer-designing-sustainable-systems-for-website-development/

GitHub (2019). "About." GitHub, Inc. https://github.com/about

Git Project, The (2019). "git." The Git Project. https://git-scm.com/

International Image Interoperability Framework (2019). "International Image Interoperability Framework." IIIF Consortium. https://iiif.io/Mauro, Aaron, et al. (2017) "Towards a Seamful Design of Networked Knowledge: Practical Pedagogies in Collaborative Teams." Digital Humanities Quarterly 11:3. http://www.digitalhumanities.org/dhq/vol/11/3/000322/000322.html

Skinner, John Skiles (2019). "love-and-seduction." Glitch. https://love-and-seduction.glitch.me/

Smithies, James, et al. (2019). "Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab." Digital Humanities Quarterly 13:1. http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html

Text Encoding Initiative (2019). "TEI: Text Encoding Initiative", Text Encoding Initiative. https://tei-c.org/

Walt Whitman Archive, The (2019). "The Walt Whitman Archive." Center for Digital Research in the Humanities. https://whitmanarchive.org/

Willa Cather Archive, The (2019). "The Willa Cather Archive." Center for Digital Research in the Humanities. https://cather.unl.edu/

Who Teaches When We Teach DH? Some Answers and More Questions.

Brian Croxall
brian.croxall@byu.edu
Brigham Young University, United States of America

Diane K. Jakacki
dkj004@bucknell.edu
Bucknell University, United States of America

Keywords: survey, pedagogy, training, curriculum, language
Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, ethnographic analysis, meta-criticism (reflections on digital humanities and humanities computing), Education/ pedagogy, Humanities computing

At the 2019 Digital Humanities Conference in Utrecht, the authors launched a survey designed to answer the question "Who Teaches When We Teach DH?" In this short talk, we will present and discuss some of the findings that may lead to a better understanding of who teaches and how DH is taught.

Constructing and Deconstructing Digital Serpentary: Goals Established, Challenges Encountered, Lessons Learned, Discoveries Made.

Alexei Alexeev
Alexei.Alexeev@uOttawa.ca
University of Ottawa, Canada

Keywords: serpent, iconography, database, metadata, collaboration
Topics: Asia, Africa, Europe, English, BCE-4th Century, database creation, management, and analysis, public humanities collaborations and methods, Art history, Galleries and museum studies

What are the best practical ways of building a specialized iconography database and developing meaningful metadata for a corpus of images selected by a common subject? While information studies offer theoretical foundations and abstract schemes, practice-based resources development such as that described in this paper provides practical outcomes and key solutions.

Serpentarium Mundi (www.serpentarium.org) is a specialized digital iconography compendium dedicated to snakes/serpents and their manifold derivatives (e.g. dragons). It catalogues representations of real and imaginary ophiomorphs in the visual arts of Old World civilizations from 3000 BC to 650 AD. The compendium's compositional framework uses the structural metaphor of traditional (print) media and consists of volumes (e.g. Coins), chapters (e.g. Deities), and articles (e.g. Agathodaemon). A minimalist approach is currently adopted: the compendium articles focus on iconographical (visual) records with the textual component limited to brief descriptions. The advanced version will include three common components: (1) a complete catalogue of the subjects' iconographic types, consisting of figures supplemented by detailed descriptions and references; (2) observations on the subject's reception in antiquity and a survey of related primary sources; and (3) the subject's interpretation in broader cultural and historical contexts and a survey of the secondary literature.

The project website serves as a limited-functionality test prototype of the SOLID (Specialized On-Line Iconography Database) digital framework. The project is a collaborative, multidisciplinary enterprise, enthusiastically sharing the values of the open data movement. It employs expertise in the fields of classics, religious studies, comparative mythology, ethnozoology, art history, and computer science. Serpentarium Mundi draws extensively upon the successes and failures of major online iconography projects, such as Iconclass, Digital LIMC, and The Warburg Institute Iconographic Database.

The presentation will inform regarding the current state of the relevant scholarship and will address several project aspects and outcomes: (1) goals established (e.g. becoming the ultimate ophidian iconography resource); (2) challenges encountered (e.g. current lack of institutional support and funding; project curator's limited technical expertise); (3) lessons learned (e.g. need to anticipate the unforeseen complexities of the evolving image classification taxonomies); and (4) discoveries made (e.g. positive identification of the nebulous iconographic subject). The following technical and organizational features will be discussed: information architecture, functionality, content selection, visual presentation, analytics, and collaboration.

The snake/serpent is a ubiquitous and intrinsically polysemantic artistic subject. The notion of "serpent" is the product of the cultural reception of the snake as a peculiar biological species, followed by its

creative modification. This transformation is usually accompanied by the complex processes of intra- and inter-cultural interpretation, appropriation, adaptation, preservation, and dissemination. Snakes and serpents play prominent roles in mythology, religion, and art and are important subjects in cultural and cross-cultural (comparative) studies (Mundkur 1983). A substantial body of literature is dedicated to the mythology, symbolism, and iconography of the serpent/dragon in diverse cultural contexts - from the Greco-Roman World and Egypt to India and China, and from the Biblical tradition to early Islam (Charlesworth 2010, Ogden 2013).

The theoretical framework of the project is informed by the pertinent discussion on iconography/ iconology (Mitchell 1986; van Straten 1994), visual semiotics (Lorenz 2016), metadata (Baca 2016), and knowledge organization (Hjørland 2008). The modified methodological scheme iconography (analytic, proto-iconic stage: pictographic inventorization) → iconology (synthetic, proper-iconic stage: descriptive identification) → iconosophy (hermeneutic, meta-iconic stage: contextual interpretation) will be introduced (cf. Panofsky 1939).

The project is inspired by the pioneering works on the ontology-based iconography conducted at the Warburg Institute (Gartner 2019) and the "Sofia School" (Pavlova-Draganova, Paneva-Marinova, and Pavlov 2011). In this particular context, ontology is understood as a knowledge-organizing process. It provides a conceptual model for describing a system (within a certain domain of discourse) consisting of types of entities, properties, and relations. The primary purpose of ontology is to limit complexity and help to convert data into information and knowledge. The use of ontologies in knowledge-based systems facilitates effective communication between various groups and domains and thus provide s the foundation for knowledge dissemination and interdisciplinary collaboration.

In this project, classification is the crucial component of the knowledge-organizing process (Szostak 2014). To facilitate the development of classification taxonomies, a novel analytical device, the three-faceted formal semantic indicator (identifier-descriptor-operator) was developed (Figure 1). The device is intended to contribute to future ontology-based iconography research as an instrument facilitating the classification, contextualization, and interpretation of images (Shatford 1986). It could also be potentially incorporated as a functional component into computational procedures involving the machine-readable ontologies (Allemang and Hendler 2011).

Main ontological categories	Formal semantic indicator (facets)	Main iconographic categories
being	identifier	⊡ deity, ⊞ hero, ⊠ mortal, ⊟ animal, ◫ creature, □ object
property	descriptor	⊛ morphology, ⊙ aspect, ⊜ prosopon, ⊖ representation ⊖ differential, ⓘ syncretism
relation	operator	⊕ association, ⊗ conflict, ⊖ context, ⊜ contiguity, ⊘ turnover

Figure 1: Formal semantic indicator and relevant ontological/iconographical categories.

The experimental application of formal semantic indicators for the classification of numismatic iconographic types will be illustrated on the example of one divine lineage, focusing on three successive generations of principal deities (Figure 2).

Figure 2a. Apollo (god of light, divination, music, and archery) confronting Python. Silver stater (reverse), Croton, Bruttium, Middle/Late Classical, 425-350 BC, © Staatliche Museen, Berlin.

Identifier:
⊡ **deity.1:** Apollo

Descriptors:
⊙ **aspect.1:** "deity.1" {holding in left hand} "association.1"
⊙ **aspect.2:** "deity.1" {holding in right hand} "association.2"
⊙ **aspect.3:** "deity.1" {targeting with "association.1" and "association.2"} "conflict.1"

Operators:
⊕ **association.1:** bow (small) {held in left hand by "deity.1" for targeting} "conflict.1"
⊕ **association.2:** arrow (small) {held in right hand by "deity.1" for targeting} "conflict.1"
⊗ **conflict.1:** serpent (Python ≅? Delphyne) (large) {swirling, rearing}; {targeted with "association.1" and "association.2" by} "deity.1"
⊖ **context.1:** tripod (large) {adorned with} "context.2"; {standing between} "deity.1" and "conflict.1"
⊖ **context.2:** garlands (two) {hanging vertically on} "context.1"

Figure 2b. Asclepios (god of medicine, son of Apollo) riding a serpent. Bronze coin (reverse), Pautalia, Thrace, Roman Imperial, Caracalla, 198-217 AD, © Coin Archives (private collection).

Identifier:
⊡ **deity.1:** Asclepios

Descriptors:
⊙ **aspect.1:** "deity.1" {riding} "association.1"
⊙ **aspect.2:** "deity.1" {holding in left hand} "association.2"
⊙ **aspect.3:** "deity.1" {facing} "association.3"

Operators:
⊕ **association.1:** serpent (large, crested, bearded, winged, fish-tailed) {coiling, rearing, *flying*}; {ridden by} "deity.1"
⊕ **association.2:** rod ≡ wand {entwined counter-clockwise by} "association.3"; {held in left hand by} "deity.1"
⊕ **association.3:** serpent (small) {rearing}; {entwining counter-clockwise} "association.2"; {facing} "deity.1"

Figure 2c. Hygieia (goddess of health, daughter of Asclepios) feeding a serpent. Bronze coin (reverse), Veria, Macedonia, Roman Imperial, Gordian III, 238-244 AD, © Staatliche Museen, Berlin.

Identifier:
⊡ **deity.1:** Hygieia ≅? Olympias

Descriptors:
⊙ **aspect.1:** "deity.1" {holding in right hand} "association.1"
⊙ **aspect.2:** "deity.1" {feeding from "association.1"} "association.2"
⊙ **aspect.3:** "deity.1" {seated upon} "association.3"

Operators:
⊕ **association.1:** phiale ≡ patera {held in right hand by "deity.1" for feeding} "association.2"
⊕ **association.2:** serpent (large) {undulating, rearing}; {fed from "association.1" by} "deity.1"
⊕ **association.3:** throne ≡ chair {adorned} {seated upon by} "deity.1"

Figure 2: Depictions of deities interacting with serpents described with the formal semantic indicators.

The presentation intends to initiate discussion, solicit feedback, and obtain expert advice from specialists in the fields of database development, cross-domain metadata, semantic web, content/network analysis, data visualization, geographic information systems, and digital museology.

Bibliography

Allemang, Dean and James Hendler. (2011). Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL, 2nd edition. Waltham: Morgan Kaufmann Publishers.

Baca, Murtha. (2016). Introduction to Metadata, 3rd edition. Los Angeles: The Getty Research Institute.

Charlesworth, James H. (2010). The Good and Evil Serpent: How a Universal Symbol Became Christianized. New Haven: Yale University Press.

Gartner, Richard. (2019). Towards an Ontology-Based Iconography. Digital Scholarship in the Humanities, fqz009, https://doi.org/10.1093/llc/fqz009 (Published: 11 February 2019).

Hjørland, Birger. (2008). What is Knowledge Organization (KO) Knowledge Organization, Volume 35, Issue 2-3, 2008, pp. 86-101.

Lorenz, Katharina. (2016). Ancient Mythological Images and Their Interpretation: An Introduction to Iconology, Semiotics, and Image Studies in Classical Art History. Cambridge: Cambridge University Press.

Mitchell, W.J.T. (1986). Iconology: Image, Text, Ideology. Chicago: The University of Chicago Press.

Mundkur, Balaji. (1983). The Cult of the Serpent: An Interdisciplinary Survey of Its Manifestations and Origins. Albany: State University of New York Press.

Ogden, Daniel. (2013). Drakon: Dragon Myth and Serpent Cult in the Greek and Roman Worlds. Oxford: Oxford University Press.

Panofsky, Erwin. (1939). Studies in Iconology: Humanistic Themes in the Art of the Renaissance. Oxford: Oxford University Press.

Pavlova-Draganova, Lilia, Desislava Paneva-Marinova, and Radoslav Pavlov. (2011). Ontological Presentation of East-Christian Iconographical Art Domain. Serdica Journal of Computing, Volume 5, Number 2, 2011, pp. 101-114.

Shatford, Sara. (1986). Analyzing the Subject of a Picture: A Theoretical Approach. Cataloging & Classification Quarterly, Volume 6, Issue 3, 1984, pp. 39-62.

Szostak, Rick. (2014). Classifying the Humanities. Knowledge Organization, Volume 41, Issue 4, 2014, pp. 263-275.

van Straten, Roelof. (1994). An Introduction to Iconography: Symbols, Allusions and Meaning in the Visual Arts. London: Taylor & Francis.

Quantifying the Unknown: How Many Manuscripts of Sévigné Still Exist?

Simon Gabay
simon.gabay@unine.ch
Université de Neuchâtel (Switzerland)

Lucie Rondeau du Noyer
lucie.rondeau.du.noyer@chartes.psl.eu
Lycée Descartes, Antony (France)

Matthias Gille Levenson
matthias.gille-levenson@ens-lyon.fr
Ecole normale supérieure de Lyon

Ljudmila Petkovic
Université de Neuchâtel (Switzerland)

Alexandre Bartz
Ecole nationale des chartes

Keywords: manuscript sales catalogues, 19th c. France
Topics: Europe, English, 15th-17th Century, 19th Century, database creation, management, and analysis, manuscripts description, representation, and analysis, Book and print history, Philology

Manuscripts can be burned, lost, forgotten, thrown away... If scholars have already tried to measure the proportion that has survived since the apparition of moveable types with Gutenberg [Weitzman, 1987], such percentages do not help editors of texts to answer a more practical question: how many documents of a given author still exist, and among them how many are accessible to scholars?

In the present paper, we want to use Madame de Sévigné (1626-1696) as a test case to calculate how many autograph manuscripts (AM) are still circulating on the market, and therefore assess precisely what is inaccessible because it is held in private collections by combining three different sources of information. First, a list of French AM held in libraries - which has been created for the occasion, because there is no catalogu e for French literature such as the one by P. Beal [2005]. Second, a list of Sévigné's AM held in historical private collections, drawn from Duchêne's edition [Sévigné, 1972-1978]. Third, a list of manuscripts drawn from fixed-price and auction catalogues [Bodin, 2000], which contains the description of hundreds of thousands of manuscripts sold over decades (fig. 1).

1. Quantification

Existing manuscripts (E) of an author (a) are either kept accessible in libraries and archives (L), either (more or less) hidden in private collections (C).

$$E_a = L_a \cup C_a$$

The problem is that while we know L, for which we have catalogues, C is unknown. To know what it represents, we can divide it in two: on the one hand we have historical collections (H), usually inherited by old families, well documented and extremely static, and on the other hand there is an unknown amount of documents circulating between private collectors (P).

$$C_a = H_a \cup P_a$$

If we cannot know P, we can use a proxy: we can deduce what is still on the market (M) by subtracting what is owned (because it has been bought) by libraries (L) from everything that has been sold (S).

$$M_a = S_a \backslash L_a$$

Because buyers are constantly intervening on the market (S), any value is only true at time t - i.e. the date of the last sale catalogue taken into account, assuming that all the previous ones have been analysed.

$$C_a(t) \approx H_a(t) \cup M_a(t)$$

With all these information, we can now deduce how many manuscripts still exist (E) if we know S.

$$E_a(t) \approx L_a(t) \sqcup H_a(t) \sqcup M_a(t) \approx L_a(t) \cup H_a(t) \cup S_a(t)$$

2 Extraction and structuring

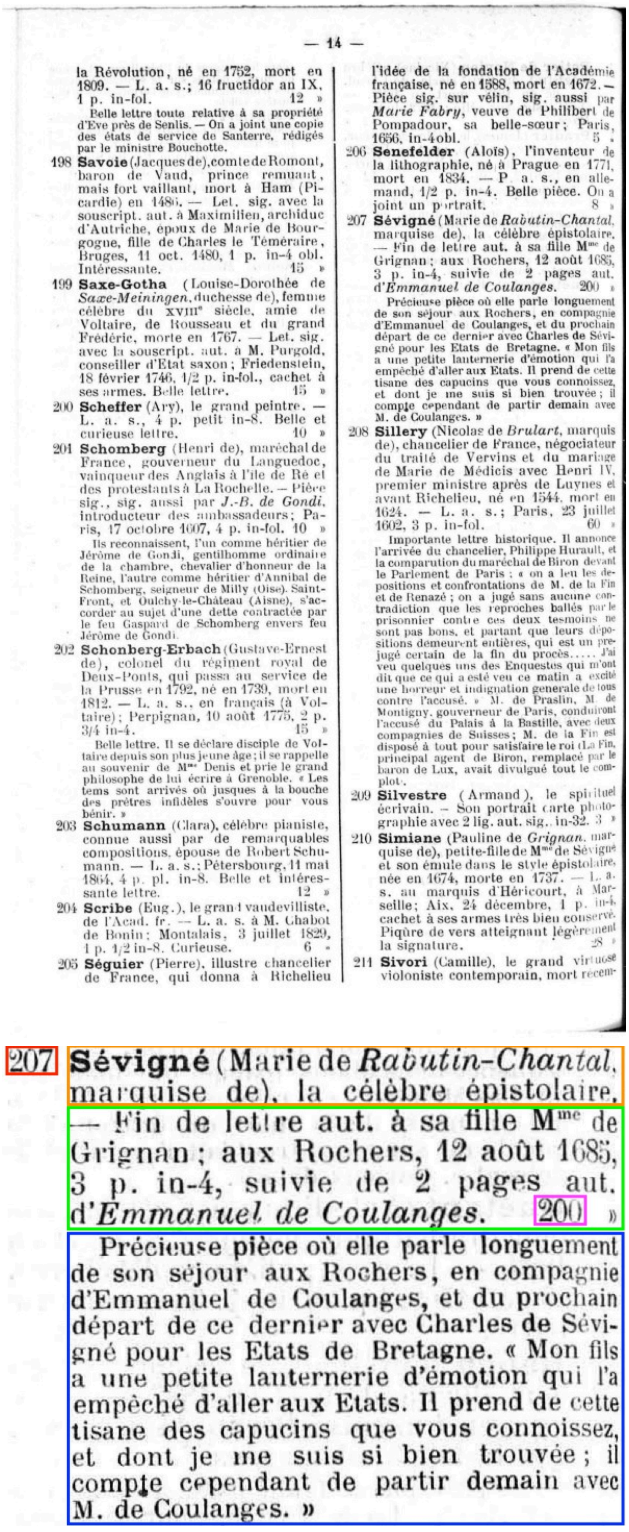


Figure 1: RDA, May 1894 (N°166)

Looking for French AM, we have concentrated our efforts on documents sold in Paris, and for financial and

time reasons, we have focused on catalogues published before c. 1900. We have retro-converted:
250 fixed-price catalogues of the Revue des autographes (RDA, cf. fig. 1).300 fixed-price catalogues of the Lettres autographes et documents (LAD).100 auction catalogues.

Because of similarities between such catalogues and dictionaries, we have been able to use GROBID dictionaries [Khemakhem et al., 2018] to process the images and transform them into a fully TEI-conformant semantic encoding (fig. 2) thanks to a custom workflow [Gabay et al., 2020].

```
<item n="207" xml:id="CAT_000256_e207">
  <num>207</num>
  <name>Sévigné (Marie de Rabutin-Chantal, marquise de),</name>
  <trait>
       <p>la célèbre épistolaire. -</p>
  </trait>
  <desc>Fin de lettre aut. à sa fille Mme de Grignan; aux Rochers,
       12 août 1685, 3 p. in-4, suivie de 2 pages aut. d'Emmanuel de
       Coulanges. 200 »</desc>
  <note>Précieuse pièce où elle parle longuement de son séjour aux
       Rochers, en compagnie d'Emmanuel de Coulanges, et du prochin
       départ de ce dernier avec Charles de Sévigné pour les Etats
       de Bretagne. «Mon fils a une petite lanternerie d'émotion qui
       l'a empêché d'aller aux Etats. Il prend de cette tisane des
       capucins que vous connoissez, et dont je me suis si bien trouvée;
       il compte cependant de partir demain avec M. de Coulanges.»</note>
</item>
```

Figure 2: XML-TEI encoding of an entry

The workflow keeps undergoing constant improvements (e.g. Rondeau Du Noyer et al. [2019]), which have led to the creation of a dedicated tool for catalogues [Khemakhem et al., 2020]. In its last version, on top of traditional features for information extraction (special characters, position on the page... in red in fig. 3), we now use typographical information (bold, italics, size of the font... in blue in fig. 3) for more precise results.

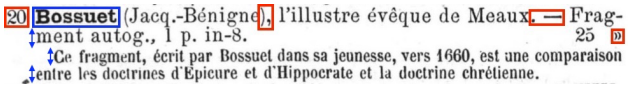

20 **Bossuet** (Jacq.-Bénigne), l'illustre évêque de Meaux. — Fragment autog., 1 p. in-8. 25 »
Ce fragment, écrit par Bossuet dans sa jeunesse, vers 1660, est une comparaison entre les doctrines d'Épicure et d'Hippocrate et la doctrine chrétienne.

Figure 3: RDA, May 1873 (N°37)

3 Annotation and reconciliation

The letter previously mentioned is not the only one of Sévigné sold during the 19th c., and it has not been sold only once:

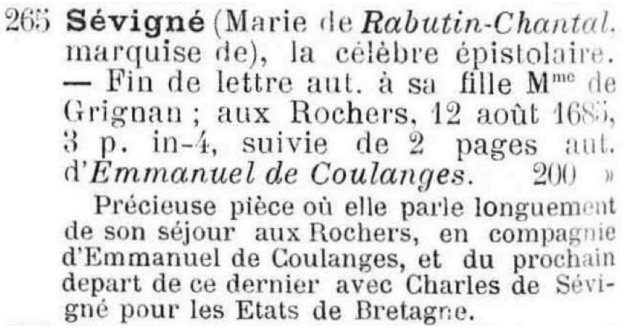

265 **Sévigné** (Marie de *Rabutin-Chantal*, marquise de), la célèbre épistolaire. — Fin de lettre aut. à sa fille Mme de Grignan ; aux Rochers, 12 août 1685, 3 p. in-4, suivie de 2 pages aut. d'*Emmanuel de Coulanges*. 200 »
Précieuse pièce où elle parle longuement de son séjour aux Rochers, en compagnie d'Emmanuel de Coulanges, et du prochain depart de ce dernier avec Charles de Sévigné pour les Etats de Bretagne.

Figure 4: RDA, July 1897 (N°200)

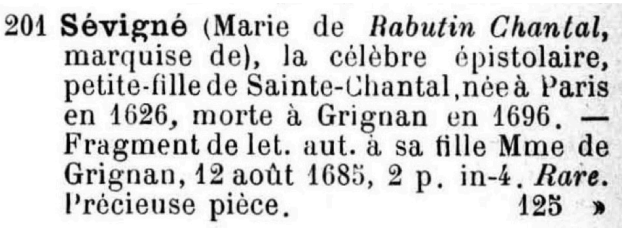

201 **Sévigné** (Marie de *Rabutin Chantal*, marquise de), la célèbre épistolaire, petite-fille de Sainte-Chantal, née à Paris en 1626, morte à Grignan en 1696. — Fragment de let. aut. à sa fille Mme de Grignan, 12 août 1685, 2 p. in-4. *Rare.* Précieuse pièce. 125 »

Figure 5: RDA, April 1902 (N°257)

Because the same item can be sold multiple times, it is crucial to transform the list obtained with the digitisation of sale catalogues into a set of unique types (or classes, cf. blue and red boxes in fig. 6), prior to comparing these types with existing documents held in libraries. Doing so, we can identify AM that have never appeared on the market (in pink and in black), document the history of those that are now in library collections (in blue and in orange) or identify "ghost" manuscripts that are still circulating on the private market (in green and red).

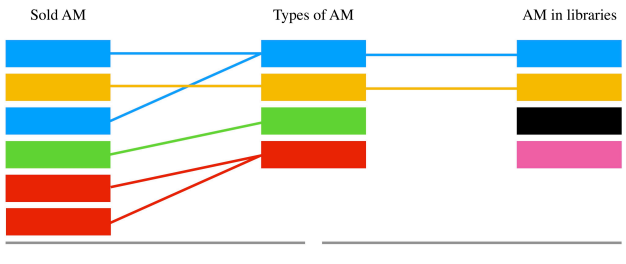


Figure 6: Reconciliation-identification process

To carry out this task, more information is required than those provided by GROBID dictionaries, we have therefore added an extra layer of information, including the type of document (L.a.s. for autograph letter signed, D.s. signed document...), its length (number of pages or folios...), its format (in-octavo, in-quarto...), its date or its price. Since these information follow either an extremely strict (1 p., L.a.s....), either a fairly common

pattern (12 janvier 1798, 19 sep. 1820...), they are tagged with regexes and dedicated python libraries in order to obtain a more fine-grained encoding:

```
<desc><term>Fin de lettre aut.</term> à sa fille Mme de Grignan; au
    Rochers, <date>12 août 1685</date>, <measure type="length">3 p.
    </measure><measure type="format">in-4</measure>, suivie de 2 pages
    aut. d'Emmanuel de Coulanges. <measure type="price">200</measure>
»</desc>
```

Figure 7: Annotated

Combined with the name of the author, such information provide a unique combination of features that can be used to compare sold documents over time, and identify not only same AM sold twice, but also different fragments of a single manuscript (tab. 1), which share part of the information only (same date, same format but different length).

RDA, May 1894 (N°166)	RDA, July 1897 (N°200)	RDA, April 1902 (N°257)
Sévigné	Sévigné	Sévigné
Fin de lettre aut.	Fin de lettre aut.	Fragment de let. aut.
12 août 1685	12 août 1685	12 août 1685
3 p.	3 p.	2 p.
in-4	in-4	in-4
200 francs	200 francs	125 francs

Table 1: Key information of three sold items from catalogues

Because we have catalogued all the known manuscripts of the marquise de Sévigné after extensive research in European and American libraries, it is possible to reconstitute part of their history thanks to the sale catalogues.

Princeton, Rare Books and Special Collections, C0710, vol. 4, f°57/N°4	Paris, BNF, NAF 717
Sévigné	Sévigné
Fin de lettre aut.	Fragment de let. aut.
12 août 1685	12 août 1685
3 p.	2 p.
in-4	in-4

Table 2: Key information of two manuscripts

4 Results

We can now offer some results:
63 sales have been identified until 190346 AM being sold at least one time, 14 at least two times13 letters out of the 46 sold are not in public libraries or archives

Following these numbers, we can say that:
c. 1% of the 1,350 letters identified by Duchêne [Sévigné, 1972-1978] are still circulating on the market.c. 5% of the total has survived but is inaccessible to scholars, if we add the 62 letters still held in the private collections of the Guitaut family in Burgundy.

Such numbers, obviously, need to be taken with caution for two main reasons. On the one hand, the oldest catalogues are not precise enough to identify exactly which AM is sold. On the other hand, the market in the 19th c. is already international, and manuscripts sold outside of France are not taken into account by our study. This second problem should receive all our attention in a near future to contribute to the history of objects [Courtin, 2020], and especially the migration of manuscripts [Burrows et al., 2019].

Acknowledgements

Many thanks to Agathe Decaster for her (crucial) help with the mathematical formulas, and her brother Erwan.

References

P. Beal. Catalogue of English Literary Manuscripts 1450-1700, 2005. URL https://celm-ms.org.uk/. https://celm-ms.org.uk.

T. Bodin. Les grandes collections de manuscrits littéraires. In A. Charon and E. Parinet, editors, Les Ventes de livres et leurs catalogues: XVIIe-XXe siècle, number 5 in Études et rencontres, pages 169-190. École des chartes, Paris, 2000.

T. Burrows, E. Hyvönen, L. Ransom, and H. Wijsman. Mapping manuscript migrations: Digging into data for the history and provenance of medieval and renaissance manuscripts. Manuscript Studies, 3(1), May 2019. URL https://repository.upenn.edu/mss_sims/vol3/iss1/13.

A. Courtin. Retour sur la datavisualisation « Sur la pistes des ventes d'antiques ». Numérique et recherche en histoire de l'art, 2020. URL https://numrha.hypotheses.org/743.

S. Gabay, L. Rondeau Du Noyer, and M. Khemakhem. Selling autograph manuscripts in 19th c. Paris: digitising the Revue des Autographes. In Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica, Quaderni di Umanistica Digitale, pages 113-118, Milan, Italy, 2020. Associazione per l'Informatica Umanistica e la Cultura Digitale. https://hal.archives-ouvertes.fr/hal-02388407.

M. Khemakhem, A. Herold, and L. Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. In GLOBALEX workshop at LREC 2018, Miyazaki, Japan, May 2018. URL https://hal.archives-ouvertes.fr/hal-01708137.

M. Khemakhem, S. Gabay, B. Joyeux-Prunel, L. Romary, L. Saint-Raymond, and L. Rondeau Du Noyer. Information Extraction Workflow for Digitised Entry-based Documents. In DARIAH Annual event 2020, Zagreb, Croatia, 2020. Digital Research Infrastructure for the Arts and Humanities. URL https://hal.archives-ouvertes.fr/hal-02508549.

L. Rondeau Du Noyer, S. Gabay, M. Khemakhem, and L. Romary. Scaling up Automatic Structuring of Manuscript Sales Catalogues. In What is text, really? TEI and beyond, Graz, Austria, Sept. 2019. URL https://hal.inria.fr/hal-02272962. https://hal.inria.fr/hal-02272962.

Sévigné. Correspondance. Gallimard, Paris, 1972-1978. ed. by R. Duchêne.

M. P. Weitzman. The Evolution of Manuscript Traditions. Journal of the Royal Statistical Society. Series A (General), 150(4):287-308, 1987. ISSN 0035-9238. doi: 10.2307/2982040. https://www.jstor.org/stable/2982040.

Widening access to Linked Ancient World Data

Sarah Middle

sarah.middle@open.ac.uk
Open University, United Kingdom

Keywords: linked data, open data, ancient world, user experience
Topics: Europe, English, BCE-4th Century, Contemporary, linked (open) data, user experience design and analysis, Archaeology, History

Introduction

Linked Data technologies connect disparate open datasets, revealing intersections between them, and could therefore be extremely powerful for Humanities research. There is much discussion among the Digital Humanities community about the production of such resources, but:

1. How are they experienced by end users?
2. How might they be made more accessible to a wider public?

My paper aims to address these questions, based on findings from recent user research. I begin by outlining my study in relation to existing research, before discussing my findings (1). I end by summarising and providing initial recommendations (2), highlighting that data openness and accuracy do not guarantee accessibility without human intervention.

Background: Linked Data and user research

Linked Data has the potential to transform Humanities research by making new connections, visualising relationships, and facilitating access to a broader range of materials. Previous research has, however, identified a number of barriers to its production in a Humanities context, including the amount of training required, lack of awareness of its benefits, and unwillingness to depart from familiar tabular formats (Barbera, 2013, pp. 96, 98; Isaksen, 2011, pp. 153-154; van Hooland & Verborgh, 2014, p. 51).

I aim to identify and address barriers to Linked Data use, rendering supposedly 'open' data resources unusable outside a relatively niche segment of the research community. Although initiatives such as Pelagios[1] emphasise the importance of a user-centred approach to development (Simon, Barker, Isaksen, & de Soto Cañamares, 2015, pp. 54-56), in general there is less existing research on Linked Data use, as opposed to production. One exception (Angelis et al., 2015) studied use of the Europeana platform; findings included identifying the need for human-curated content as a 'way in' to the data.

Limiting my study to researchers of the Ancient World, a microcosm of the Humanities, I set out to identify which Linked Data resources are used most frequently, explore users' experiences of these resources, and recommend how they might attract a wider audience.

Methodology: survey and interviews

To obtain a breadth of views and in-depth understanding of specific cases, the study included a survey and interviews. Based on the assumption that few Ancient World researchers have used Linked Data, but many could benefit from it, I aimed the survey at all Ancient World researchers who use digital tools and resources, intending to apply more generic findings to Linked Data specifically.

212 participants responded, who were predominantly:

- associated with universities (academics, researchers, or students);
- aged 25-44;
- from the UK, US or Germany (although 32 countries were represented).

There were slightly more women than men, with a small proportion who preferred not to provide their gender, or to self-describe. I selected an interview sample of 17, ensuring differing levels of technical experience, while being broadly demographically representative of the survey population. The combined survey and interview results comprise a cross-section of views relating to Linked Ancient World Data at the present time.

Findings: curation and documentation

The 23% of participants who had knowingly used Linked Data were asked about the tool or resource with which they are most familiar. Those discussed most often were Pleiades[2], Pelagios and Papyri.info[3]. Features of these resources that participants particularly appreciated included:

- accessing multiple data sources from a single point;
- advanced search features;
- viewing connections between related entities.

Such features were also of interest to Ancient World researchers more generally. Even those with limited technical experience showed an interest in integrating multiple resources and incorporating advanced query options to improve efficiency of discovery. Linked Data resources therefore clearly have a broad potential appeal.

User experience comments on Pleiades, Pelagios and Papyri.info were largely positive, for example:

"*I found it user-friendly and intuitive*." [PART007, Survey] - on Pelagios

However, some participants identified potential barriers:

- it was sometimes unclear where to start;
- documentation could be improved.

Subsequent interview discussions indicated that, although clear documentation was appreciated, less technically experienced participants were unlikely to consult it at all due to an assumption that it would be too complex to understand. Participants' suggestions for resolving the above barriers included:

- developing curated collections;
- incorporating use cases ("Clear case studies for non-technical people." [PART054, Survey]).

Both these suggestions harness human skills rather than relying solely on richness of data.

Conclusions: the human factor

Findings from this study illustrate the importance of human intervention to ensure that Linked Open Data is usable and accessible by a wider public, rather than leaving the data to speak for itself. Specific recommendations include:

- Curating content to provide a starting point for new users.
- Improving documentation, by:
    - ensuring clarity of language for less technically experienced users;
    - providing examples of how the resource has been used in the past or might be used in future;
    - integrating more fully with the user interface, e.g. signposting relevant sections at key points in the user journey.

Although any updates to a digital tool or resource require an investment of time and funds, the above actions would be less costly than redesigning a user interface and could potentially be as effective in broadening the potential audience. While the focus of this study was Linked Data, the findings should be viewed in the context of participants' wider digital experiences; as such, these recommendations could equally apply to resources that employ a different technological approach.

This work forms part of a wider PhD study, aiming to identify where Linked Data might be most effectively integrated with Ancient World research methodologies. The resulting recommendations should enable researchers with varying levels of technical skill to unlock its full potential.

References

Angelis, S., Benardou, A., Chatzidiakou, N., Constantopoulos, P., Dallas, C., Dunning, A., ... Waterman, K. (2015). Europeana Cloud Deliverable 1.3 User Requirements Analysis and Case Studies Report. Content Strategy Report. Retrieved from Europeana website: http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Cloud/Deliverables/D1.3%20D1.6%20User%20Requirements%20Analysis%20and%20Case%20Studies%20Report%20Content%20Strategy%20Report.pdf

Barbera, M. (2013). Linked (open) data at web scale: Research, social and engineering challenges in the digital humanities. JLIS.It, 4(1), 91-101.

Isaksen, L. (2011). Archaeology and the semantic web (University of Southampton). Retrieved from http://eprints.soton.ac.uk/206421

Simon, R., Barker, E., Isaksen, L., & de Soto Cañamares, P. (2015). Linking Early Geospatial Documents, One Place at a Time: Annotation of Geographic Documents with Recogito. E-Perimetron, 10(2), 49-59.

van Hooland, S., & Verborgh, R. (2014). Linked data for libraries, archives and museums: How to clean, link and publish your metadata. London: Facet Publishing.

[1] https://pelagios.org/

[2] https://pleiades.stoa.org/

[3] http://papyri.info/

The Miiyupimatisiiun Research Data Archives Project: Co-establishing a usable Indigenous data management system.

Samuel Mickelson
samuel.mickelson@ryerson.ca
Ryerson University, Canada

Keywords: Digital archives, Digitization, Indigenous archives, Indigneous data sovereignty

Topics: English, North America, 20th Century, Contemporary, digital access, privacy, and ethics analysis, digitization (2D & 3D), Anthropology, Library & information science

The Miiyupimatisiiun Research Data Archives Project (MRDAP) is a collaborative digitization and data transfer initiative between medical anthropologist Dr. Naomi Adelson and Whapmagoostui First Nation (FN) in the territory of Eeyou Istchee (James Bay). The project is the result of a longstanding research partnership between Adelson and Whapmagoostui FN that dates back to the late 1980's. The research data archives consists of a large corpus of materials, including: over 60 hours of digitized audio recordings; over 2,000 digitized colour photographs; and over 50 interview transcripts which have been digitally migrated from 3 1/2 inch floppy disks.

Figure 1: Aerial view of Whapmagoostui First Nation

The MRDAP raises important questions regarding the limitations of existing library and archival principles; the legal and practical barriers to Indigenous data sovereignty; and the role of repositories with colonial histories in stewarding research data originating from Indigenous communities. Can the materials in the archives coexist as Indigenous cultural memory and academic research data? In what ways are these values (in)commensurable? How is community control over data access, dissemination, reproduction, and preservation being navigated? What ought to be the role of the non-Indigenous cultural heritage professional in this relationship?

Figure 2: Photograph from the Miiyupimatisiiun Research Data Archives

This presentation will engage with these questions from the perspective of the settler archivist who digitized the research data archives. We situate the project within the growing body of literature on Indigenous archives and sustainable digital initiatives (Byrne, 2009; Christen, 2011; McKemmish, Faulkhead, & Russell, 2011; Swanson, 2015; Whaanga et al., 2015). We also describe the ways in which key guidance documents such as the OCAP(TM) Principles and the Protocols for Native American Archival Materials have shaped project design and implementation (First Archivist Circle, 2007; First Nations Centre, 2007). In particular, we describe the ways in which these documents have influenced our decision-making regarding how the research data archives will be managed, accessed, stored, and preserved over the long-term. Overall, we argue that the project serves as a valuable case study on the development of a sustainable Indigenous data management system.

Early Modern Digital Itineraries: Modeling European Place and Space, 1545-1747

Rachel Clare Midura
rmidura@stanford.edu
Stanford University, United States of America

Keywords: network, spatial, dynamic, temporal, history
Topics: Europe, English, 15th-17th Century, 18th Century, network analysis and graphs theory and

application, spatial & spatio-temporal analysis, modeling and visualization, Book and print history, History

Before the advent of formal cartography and its emphasis on observation, accuracy, and reliance on global standards, the itinerary was the height of geographic knowledge. Lists of cities and their relative distances, represented by many national "miles" or eventually the location of postal waystations, opened European travel to a broad readership. By the mid seventeenth-century, private and public individuals had access to, and knowledge of, a wide network of mounted couriers and staging posts, connecting trade, travel, and epistolary exchange from the hubs of Milan, Venice, and Innsbruck to Constantinople, Madrid, and the New World. European postal networks were complex carrefours, bringing together merchants, pilgrims, tourists, and diplomats in both real and imagined space.

Digital spatial approaches often rely upon modern mapmaking and its assumptions of a decentered viewpoint, direct distances, and national boundaries. Scholars of indigenous mapping such as Paul Carter, Deborah Kirk and Jeremy Mikecz, have demonstrated the assumptions and limitations of such an approach. The application of Social Network Analysis (SNA) to a corpus of 70 published itinerary books models the organizing logic of the itinerary genre and hierarchization of regions, cities and routes. While the pilgrimage path of St. James and transalpine commercial routes were widely republished, dynamic networks based on the dates of first and last publication indicate the influence of new postal hubs, sea travel, and cartography on early modern conceptions of a connected Europe.

I draw on the prior work of Shawn Graham and Elijah Meeks to develop dynamic cartograms, while adding new emphasis to their documentation of a contemporary mental map. A combination of spatial representation with Social Network Analysis (SNA) better recreates the early modern experience of space. Utilizing the SNA, TSNA, and networkDynamic packages in R, I construct and measure dynamic network models for comparison with spatial maps. The itineraries linked locations along chained journeys from origin, to intermediary, to final destination, as they were experienced by travelers or couriers on the ground. The experience of travel was highly variable based on the political and environmental circumstances; a dynamic network more accurately conveys the changing nature of European space. Digital methods prove key for moving between scales of consideration, from following the fate of one city, to many linked cities, to entire regions or the network as a whole. Data will be made available on Github with the publication of a related article.

Visceral Histories: Analyzing Transmission through Katherine Dunham's Dance Repertory

Bench Harmony
bench.9@osu.edu
The Ohio State University

Kate Elswit
kate.elswit@cssd.ac.uk
University of London, Royal Central School of Speech and Drama

Keywords: Dance History, Katherine Dunham, Performance, Bodies, Diaspora
Topics: Comparative (2 or more geographical areas), English, North America, 20th Century, cultural analytics, spatial & spatio-temporal analysis, modeling and visualization, History, Performance Studies: Dance, Theatre

This paper attends to intergenerational and intercultural travel of dance-based knowledge in the diasporic practice of choreographer Katherine Dunham, in order to contend that digital methods can evidence and elaborate bodily experience in visceral histories at the crossroads of dance and digital humanities. Using data manually curated from Dunham's extensive archives, we seek to better understand dynamic processes of movement circulation, from body-to-body transmission among company members, to in- and outflows as the company toured the globe. We situate the research within the challenges of using data to attend to bodies and embodied experience. Then we turn to two key areas of data curation, visualization, and analysis to consider how diaspora moved through Dunham's works and through the bodies of performers: the embodied practices through which Dunham's performers contributed to developing and maintaining her choreography, and the ways in which the repertory itself transmits embodied knowledge across generations and cultures.

Virtual Space, Voice, and Gender in Recorded Popular Music from 2008–2018

Michèle Duguay
mduguay@gradcenter.cuny.edu
The Graduate Center, CUNY

Keywords: Virtual Space, Voice, Gender, Recorded Popular Music
Topics: English, North America, Contemporary, music and sound digitization, encoding, and analysis, Gender and sexuality studies, Musicology

I develop an approach for locating the voice in a virtual space in recorded popular music. I use the method to examine how the spatialization of the voice within this space conveys gendered meanings. The method locates the voice in virtual space along six spectra: (1) left/right; (2) high/low; (3) localized/diffuse; (4) prominent/blended; (5) flat/reverberant; and (6) single vocal track/multiple vocal tracks.

I illustrate the methodology by analyzing a corpus of 133 tracks from the 2008-18 Billboard Year-end charts. My analysis demonstrates that male singers' voices tend to be localized, prominent, and located in a flat space. Conversely, female singers' voices are more likely to be diffuse, blended with the sonic environment, treated with echo and reverberation, and layered with multiple vocal tracks. The project contributes to analytical literature on recorded music while documenting the ways in which gendered meanings are sonically constructed.

Digital History of Post-Soviet Virtual Museums

Nadezhda Povroznik
povroznik.ng@gmail.com
Perm State University, Russian Federation

Keywords: virtual museum, web archive, museum studies
Topics: Asia, Europe, English, Contemporary, cultural analytics, digital archiving, Galleries and museum studies, History

The history of virtual museums began long before the appearance of web-museums on the Internet. Its trajectory is associated with the deepening and intensification of the use of information technologies in various areas of museums. The modern information environment has reached a high level of development, and users can now have many varied experiences online in order to engage with digital cultural heritage.

It is very difficult to give a modern definition of the term "virtual museum", since an updated definition of the term "museum" is still under development by ICOM. A group of experts working on a large European VIMM project gave the following definition of the concept of "virtual museum" as "a digital entity that draws on the characteristics of a museum, in order to complement, enhance, or augment the museum through personalization, interactivity, user experience and richness of content" [1].

The digital history of virtual museums has been previously considered from different perspectives. Erkki Huhtamo's article is noteworthy as leading the early history of virtual museums from the pre-digital era and changing the way they look at exhibition design[2]. Paul L. Arthur approached the comprehension of the history of virtual museums from the point of view of the role and importance of interactivity, increasing the ability of the user to receive diverse experience in the online environment [3]. It should be noted that the history of virtual museums has been formerly examined in a global context, while this work is devoted to the peculiarities of the history of virtual museums in the post-Soviet space.

The focus of this study is virtual museums in the post-Soviet space, which includes the former Soviet republics that left the USSR after its collapse in 1991.

The main source for studying the digital history of virtual museums has become the resources of web archives. The Museum.ru resource, snapshots of which are presented in WayBack Machine [4], has become an important source of information for studying Russian museum web space. The site was created as a platform for museums, allowing each museum to create its own web-museum resource with a wide range of capabilities. The specified aggregator contains statistics on such web resources, which is presented in the diagram (Fig. 1).

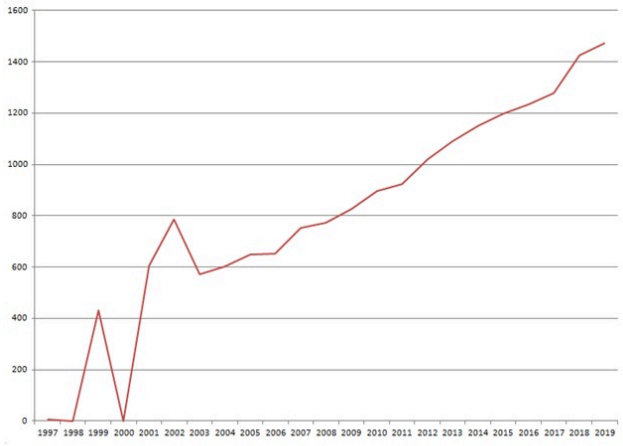


Figure 1. Russian museum web-sources in the dynamics (1997-2019), based on Museum.ru statistics

It should be noted that the initial policy of this site was to provide space on the platform to any organization claiming to be a "virtual museum". Until 2003 many private initiatives were placed there and they were not related to the GLAM industry. Since 2003, only web resources of actual existing museums of Russia have remained in the catalog (no statistics were available for 2000).

As a result of the work on the project, the following main features of the history of virtual museums of the post-Soviet space can be highlighted:

First, during the existence of the USSR, the experience of using information technologies in museums was limited to separate experiments of documenting individual collections (for example, the Hermitage developed an information system for data processing of antique art bronze and petroglyphs of Central Asia). With the collapse of the USSR, the globalization of the market in the post-Soviet space, and the widespread penetration of computers and software from Western countries, it was necessary to abandon partially the accumulated experience due to the technical incompatibility of Soviet technologies with the new foreign hardware and software. Whereas European and North-American countries actively used multimedia content, making it the central element of virtual museums online, in the post-Soviet space multimedia rarely went into the virtual space, firmly entrenched in the commercial sector with access to CDs.

Secondly, since the beginning of the Internet era, non-profit and philantropic organizations that support and stimulate the development of this sector, in particular the non-profit partnership "Museum Automation and Information Technology" (ADIT), have been significant in the development of the museum environment. Non-governmental grant organizations support the community of museum professionals in the post-Soviet space, introducing innovations in museums and developing virtual museums. Among the countries studied, Russia has one of the leading positions in the development of a virtual museum space in technological and innovation terms.

Thirdly, the representation of museum collections has gone from simple illustrations and museum narrative on exhibition topics to digital catalogs with a wide range of search possibilities, digital excursions and museum storytelling. An issue to note is the limited work with open licenses for the use of digital objects, which complicates the wider use of collections.

Fourth, searches for the digital identity of virtual museums in different countries of the post-Soviet space, as well as an analysis of the visual images used, demonstrate a certain degree of dependence on the official political discourse. This is most acute in Russian and Ukrainian virtual museums of history, which showed a change of images after the events of 2014 and the aggravation of inter-state relations.

In general, the digital history of the virtual museums of the former Soviet Union shows that the most advanced virtual museums have a long digital history, such as The Pushkin State Museum of Fine Arts. The general level of virtual museums is significantly behind those globally in the quality of digitization, the methods of storytelling in web museums, and the creation of an audience-oriented digital content. At the same time, history shows that it is possible to adopt the positive experience of leading museums for effective further development.

References:

[1] Virtual Multimodal Museum. ViMM Plus. URL: https://www.vi-mm.eu/2018/01/10/the-vimm-definition-of-a-virtual-museum, last accessed: 10/06/2020

[2] Huhtamo, E. (2002) On the Origins of the Virtual Museum, Nobel Symposium "Virtual Museums and Public Understanding of Science and Culture" May 26-29, 2002, Stockholm.

[3] Arthur P. L. (2018). Engaging Collections and Communities: Technology and Interactivity in Museums. In Proceedings of the Digital Humanities in the Nordic Countries. 3-rd Conference, Helsinki, Finland, March 7-9, 2018. Pp. 250-262.

[4] WayBack Machine. Internet Archive. URL: https://archive.org/web/, last accessed: 10/06/2020.

Residual or Essential: On the Problem of Overlap as Argument to Broaden the Solution Space of Markup

Ian Neill
Akademie der Wissenschaften und der Literatur. Mainz, Germany.

Joris J. Van Zundert
joris.van.zundert@huygens.knaw.nl
Huygens Institute for the History of the Netherlands – Royal Netherlands Academy of Arts and Sciences. Amsterdam, The Netherlands.

Keywords: stand off, markup, annotation, multidimensional text
Topics: Europe, English, North America, Australia/Oceania, Contemporary, annotation structures, systems, and methods, text encoding and markup language creation, deployment, and analysis, Humanities computing, Philology

We return to the problem of overlap in embedded markup. We argue that the addressing of overlap as a merely technical issue signals a myopic view in the field of digital textual scholarship. Overlap is not a residual problem in an otherwise evolved and exclusive solution to text modeling. Instead, the problem of overlap indicates that digital text should be approached scholarly by application of multiple modeling paradigms. To further suport our argument we present how the SPEEDy standoff properties editor succeeds in combining different markup ontologies productively.

A Neural OCR Engine for North Saami

Andre Kåsen
National Library of Norway

Håvard Østli
National Library of Norway

Andrea M. Huus
National Library of Norway

Lars Johnsen
National Library of Norway

Keywords: OCR, Neural Networks, Tesseract, North Saami
Topics: Europe, English, Contemporary, artificial intelligence and machine learning, digitization (2D & 3D), Humanities computing, Linguistics

The DH-LAB at the National Library of Norway can announce that we have an open-source optical character recognition (OCR) engine for North Saami in construction. North Saami is an under-resourced indigenous minority language recognized by the Norwegian State. The OCR engine is induced with the system Tesseract by the means of cross-lingual model transfer. When evaluating the model on a held-out portion of the ground truth, it reaches a bag-of-words F1 measure of 0.98 %. The OCR engine in question will be the first freely available OCR engine for North Saami.

Visual Authority, Privacy, and Surveillance in the Age of Deepfakes

Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Keywords: critical data studies; software studies; algorithmic bias; surveillance; visual culture
Topics: Global, English, Contemporary, artificial intelligence and machine learning, digital access, privacy, and ethics analysis, Computer science, Media studies

In this paper I present a study of the impact of deepfake technologies in digital modes of communication. I illustrate that deepfakes make novel use of photographic documents as tools for the violation of individual privacy and in doing so offer a fundamentally different challenge to concepts of visual authority. Digital surveillance and "big data" systems monitor and record the observed actions and behaviours of people. Algorithmic methods extend

surveillance to predictions of behaviours that people may possibly, in the future engage in. I situate deepfakes as a further violation of privacy applied to actions and behaviours that have not and never will occur. By establishing deep-learning based forgeries as an extension of existing tools of digital surveillance, I aim to provide a framework for the critical analysis of a new digital technology that is well positioned to become ubiquitous in the very near future.

The Words Themselves: A Content-Based Approach to Quote Attribution

Adam Hammond
adam.hammond@utoronto.ca
University of Toronto, Canada

Krishnapriya Vishnubhotla
University of Toronto, Canada

Graeme Hirst
University of Toronto, Canada

Keywords: quote attribution, text analysis, novel, machine learning, annotation
Topics: Europe, English, North America, 19th Century, 20th Century, Contemporary, attribution studies and stylometric analysis, text mining and analysis, Computer science, Literary studies

Quote attribution is the identification of the speaker of a quotation in a given text. It requires reasoning about conversational patterns and contextual clues, and is especially complex in literary texts. We present a semi-supervised iterative classification approach to quote attribution that is based on ideas from computational stylometry, using the content of the quotation to distinguish between speakers. We achieve an accuracy of 77.3% on the QuoteLi quote attribution corpus. Despite certain limitations, we show that our method is a competitive alternative to systems based on contextual clues, and a viable complement to them.

A Conceptual Framework for the Analysis of Multilayer Networks in the Humanities

Martin Grandjean
martin.grandjean@unil.ch
University of Lausanne, Switzerland

Keywords: Social Network Analysis, Multilayer Networks, Data Visualisation, Historical Network Research, Network Analysis
Topics: Global, English, Contemporary, network analysis and graphs theory and application, spatial & spatio-temporal analysis, modeling and visualization, History

If network analysis has made its way into the humanities toolbox, and especially in history (Düring et al. 2016), it is because it helps to grasp the complexity of the objects of these disciplines. However, to understand the multidimensionality of the data requires a consequent reflection on its modeling.

This paper seeks to be part of a series of publications aimed at making advanced network analysis concepts more accessible to the humanities scholars: from ontological questions (Langmead et al. 2016) to the necessary discussion of the integration of temporality in graphs (Lemercier 2015; Conroy et al. 2019), the development of typologies of uses (Grandjean 2017a) or attempts to provide aids to interpretation (Grandjean & Jacomy 2019). The question of multilayer networks becomes especially more and more important, whether in a general way (McGee et al. 2016) or applied to the humanities (eg. Vugt 2017; Grandjean 2017b).

Our purpose is to discuss a unifying conceptual framework allowing the transition between a current formal multilayer model (Kivelä et al. 2014; Knudsen et al. 2019) and the language of the humanities. This framework is expressed by a visual representation that contains a multiplicity of layers that synthesizes and clarifies the different possible networks and facilitate the appropriation of the model by researchers.

Description of the framework

The key element of this concept is the distinction between "layers" and "levels": within a horizontal level, 1-mode graphs all have the same set (and type) of nodes (visualized with a fixed position or not, depending on

our will to represent the different structures or to help keeping the "mental map" between each other). A level is therefore a multiplex system of layers. The levels can be connected together by edges that express the relation between two types of nodes (2-mode).

It is to be noted that the formal model does not imply a hierarchy between layers. Here, this verticality is only a visual artifact to facilitate the expression of all kind of humanities datasets. Each level represents one entity type and relationships within. It can have many layers organized by "aspects" and include edges between these layers. The inter-level relations can be represented on a layer that contains a 2-mode network (which can also be deployed according to different aspects).

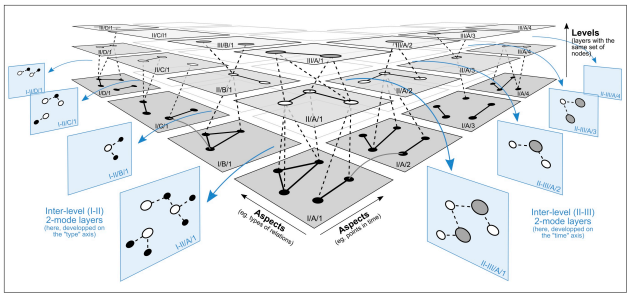


Figure 1: Multilayer network system made of three 1-mode networks organised on three layers set ("levels") connected by two 2-mode networks. At each level, this representation explicit the possibility of developing the graph depending on (at least) two aspects.

The visual expression of this framework (fig. 1) looks quite similar to an OLAP Graph Cube (Zhao et al. 2011) or a Space-Time Cube Visualization (Bach et al. 2014). The proposed framework is however not a cube: this well aligned appearance is a convenient way of representing things to make them less abstract so that they can be used to discuss modeling issues. Not all levels necessarily have the same dimensions and do not need to be aligned. Moreover, there can be more than three axes since it is an intellectual tool and not a graphic construction.

Use cases

This visual convention allows to design "scenarios" where the model coincides with data from applied research to make the layers within them explicit:
Affiliation networks (fig. 2.1) based on interlocking data with multiple companies, institutions, prosopographic categories (eg. David & Westerhuis 2014).

Character networks (fig. 2.2) based on theatre plays with addition of information on groups, families, temporality (eg. Xanthos et al. 2016).

Circulation networks (fig. 2.3) with multiple transportation means, places, goods, routes, letters (eg. Orengo & Livarda 2016).

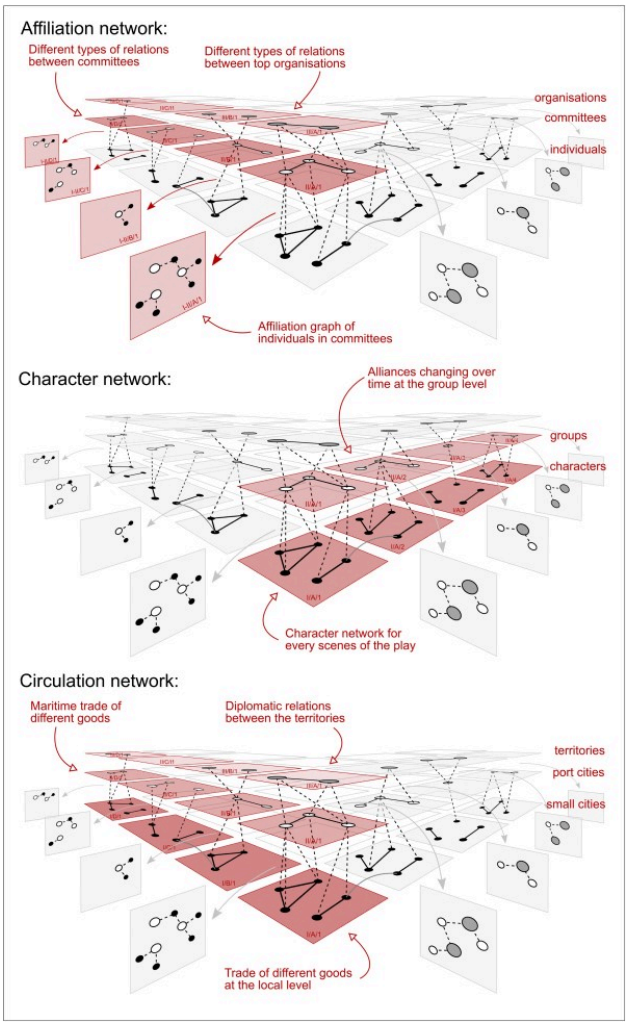


Figure 2: Modeling example of an affiliation network (1), a character network (2) and a circulation network (3).

References

Bach, B., Dragicevic, P., Archambault, D., Hurter, C., & Carpendale, S. (2014). A Review of Temporal Data Visualizations Based on Space-Time Cube Operations. In Eurographics Conference on Visualization, Swansea.

Conroy, M., Elo, K., Heyer, G., ... Weingart, S. (2019). Visualizing Networks and Temporality. Network Visualization in the Humanities (Dagstuhl Reports), 8(11), 149-150.

David, T., & Westerhuis, G. (Eds.). (2014). The Power of Corporate Networks. A Comparative and Historical Perspective, New York: Routledge.

Düring, M., Eumann, U., Stark, M., & Keyserlingk, L. (Eds.). (2016). Handbuch Historische

Netzwerkforschung: Grundlagen und Anwendungen, Berlin: LIT-Verlag.

Grandjean, M. (2017a). Analisi e visualizzazioni delle reti in storia. L'esempio della cooperazione intellettuale della Societa delle Nazioni. Memoria e Ricerca, 55(2), 371-393.

Grandjean, M. (2017b). Multimode and Multilevel: Vertical Dimension in Historical and Literary Networks. In Digital Humanities, Montreal.

Grandjean, M., & Jacomy, M. (2019). Translating Networks: Assessing Correspondence Between Network Visualisation and Analytics. In Digital Humanities 2019, Utrecht.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. Journal of Complex Networks, 2(3), 203-271.

Knudsen, S., Aerts, J., Archambault, D., ... Von Landesberger, T. (2019). Unifying the Framework of Multi-Layer Network and Visual Analytics. Visual Analytics of Multilayer Networks Across Disciplines (Dagstuhl Reports), 9(2), 19-23.

Langmead, A., Otis, J. M., Warren, C. N., Weingart, S. B., & Zilinksi, L. D. (2016). Towards Interoperable Network Ontologies for the Digital Humanities. International Journal of Humanities and Arts Computing, 10(1), 22-35.

Lemercier, C. (2015). Taking time seriously. How do we deal with change in historical networks? In M. Gamper, L. Reschke, & M. Düring, eds., Knoten und Kanten III. Soziale Netzwerkanalyse in Geschichts- und Politikforschung, Bielefeld: Transcript Verlag, 183-211.

McGee, F., During, M., & Ghoniem, M. (2016). Towards Visual Analytics of Multilayer Graphs for Digital Cultural Heritage. In IEEE Viz Workshop "Visualization for the Digital Humanities."

Orengo, H. A., & Livarda, A. (2016). The seeds of commerce: A network analysis-based approach to the Romano-British transport system. Journal of Archaeological Science, 66, 21-35.

Vugt, I. van. (2017). Using Multi-Layered Networks to Disclose Books in the Republic of Letters. Journal of Historical Network Research, 1(1), 25-51.

Xanthos, A., Pante, I., Rochat, Y., & Grandjean, M. (2016). Visualising the Dynamics of Character Networks. In Digital Humanities, Krakow, pp. 417-419.

Zhao, P., Li, X., Xin, D., & Han, J. (2011). Graph Cube: on Warehousing and OLAP Multidimensional Networks. In SIGMOD, Athens.

Analyzing Border Crossings in Contemporary Irish Literature with Literary Mapping

Kelli Maloy
kem25@pitt.edu
University of Pittsburgh at Greensburg, United States of America

Fiona Carter
frc23@pitt.edu
University of Pittsburgh at Greensburg, United States of America

Keywords: literary mapping, spatio-temporal analysis, Ireland, Irish studies, Belfast
Topics: English, North America, Contemporary, spatial & spatio-temporal analysis, modeling and visualization, text encoding and markup language creation, deployment, and analysis, Humanities computing, Literary studies

Our project analyzes works by Belfast writers Anne Devlin and Lucy Caldwell to develop a spatio-temporal analysis of the ways in which women are depicted as crossing borders in public space, particularly during the Troubles. We used GIS, and specifically the open-source geodata and software QGIS and OpenStreetMaps. To visualize the external or implied aspects of the locations identified in the works, we embedded socio-historical information, such as census data. We used OCR and XML to pull information from the texts, though copyright prohibited us from making these texts publicly available. Once we have compiled our data, one way in which we plan to consider it is through the lens of Henri Lefebvre's The Production of Space.

Graph models for the genesis of Goethe's Faust

Thorsten Vitt
thorsten.vitt@uni-wuerzburg.de

Julius-Maximilians-Universität Würzburg, Germany

Keywords: digital edition, graphs, chronology
Topics: Europe, English, 18th Century, 19th Century, Contemporary, data modeling, scholarly editing and editions development, analysis, and methods, Computer science, Literary studies

Johann Wolfgang Goethe has been working on his drama *Faust* for almost his entire life. As witnesses of his work, 556 manuscripts are currently known. Together with the 15 relevant prints that have been published during Goethe's lifetime, a newly constituted text and supplementary material have been published in a hybrid edition (Goethe 2019).

The edition aims to provide insight into the genesis of the drama's text. Yet, the order in which the individual manuscripts have been created as well as the dating of the acts of inscription are subject to more than 100 years of research and editorial activity. Most studies only deal with a handful of witnesses and provide either a relative chronology or a broad dating for them. The only attempts to aggregate individual statements in order to place all relevant objects in a chronological-stemmatic relation to date have been provided by Renate Fischer-Lamberg. Her stemmata for two acts of *Faust II* (Fischer-Lamberg 1955, 150-66) probably mark the practical limit of how much of this information can be grasped by human means alone.

Basic graph model

In order to facilitate algorithmic aggregation, the edition project has first formalized the assertions from the research literature as illustrated in fig. 1.

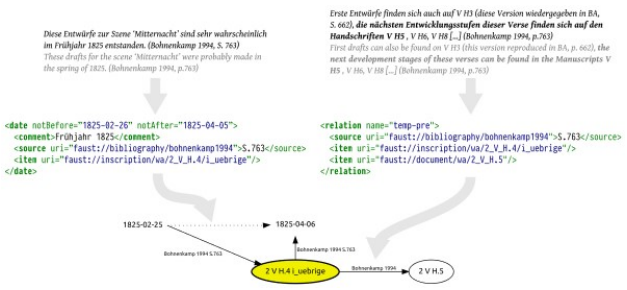


Figure 1: Formalizing absolute datings and relative chronologies from literature, and including them in a common graph model (Vitt and Brüning 2019)

Early attempts of working with the data used logic solvers and a set of rules to infer new, derived assertions (Wissenbach, Pravida, and Middell 2012). In the current approach, graph based models are used since they provide dual benefit: There is a wealth of algorithms to answer various questions on graphs, and subgraphs can be easily visualized in order to understand and justify a certain witness's dating.

The basic model combines all formalized assertions into a single directed multi-graph as illustrated in fig. 1. Under the assumptions that all assertions are correct, we can now infer an order that is consistent with all assertions (a topological sorting (Kahn 1962)), and we might be able to infer or improve limits for the absolute dating of a witness by looking for the nearest reachable dates in the graph.

Conflicts

Unfortunately, there are contradictions between some of the assertions. In the graph, contradictions appear as *cycles* and prevent both kinds of deduction outlined above. Since the data features a strongly connected component with 477 documents and 2136 edges inducing millions of cycles, it is not possible to determine manually which edges best to remove to make the graph acyclic. While this problem generally is NP-hard (Karp 1972), an algorithm by Baharev, Schichl, and Neumaier (2015) solves it for our graph, suggesting a relatively small subset of edges to remove (171 of 3480). Edge weights based on the assertions' sources influence this algorithm.

For each automatically determined conflict, a visualisation indicating conflicting assertions makes it easy to manually check against the original sources (fig. 2). The result can be fed back into the data by assigning edges a specificly large or low weight.

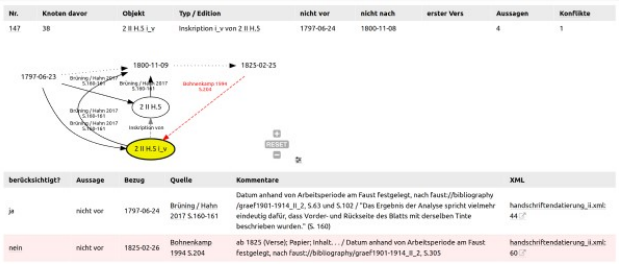


Figure 2: Conflict visualization for an automatically discarded edge (dashed red)

Inscriptions

To complicate things further, a single document may contain multiple *inscriptions* (Hoenen and Brüning 2019), i.e., it may have been written on in completely separate working phases. E.g., the single sheet 2 II H.5 (cf. fig. 2) has verses belonging to different parts of Faust on the recto and verso side, so Bohnenkamp (1994) deduced from what is known about Goethe's working phases that they might have been written 25 years apart. (Brüning and Hahn (2017) show both sides are written with the same ink, indicating synchronicity instead).

There are assertions both on inscriptions as well as on their respective witness as a whole (fig. 3), and it is not always clear or consistent between authors which parts of the text belong to which inscription.

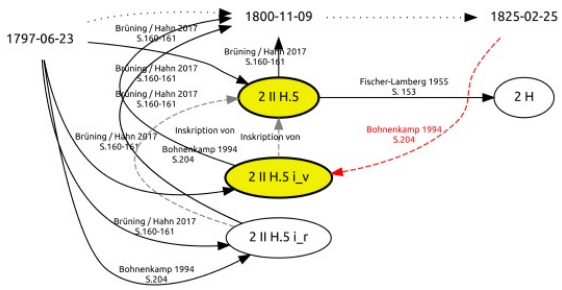


Figure 3: 2 II H.5 and its inscriptions

Model variants

There have been experiments with variants on the model to deal with these differences:

- Copying information about inscriptions to their corresponding witnesses changed the ordering of 17 datable objects.
- Research literature provides a few assertions about "approximately synchronous" witnesses. To include this information, incoming and outgoing edges have been distributed within each group of synchronous witnesses. This induced absolute datings for more nodes, and it also changed the ordering for up to 67 objects.
- Witnesses and inscriptions were modelled using two linked nodes each, representing the start and end of the working phase. All incoming edges end at the start node, all outgoing edges emerge from the end node, and inscriptions are linked to their inscription in order to happen 'during' the timespan of the witness (fig. 4). This model decreased the conflicting edges (to 139), and it has some influence on the ordering.

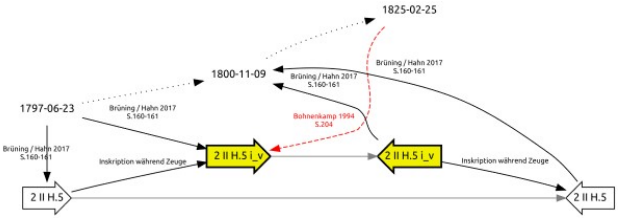


Figure 4: Working phase as interval

Evaluation

There is no pre-known correct ordering to use as a gold standard for evaluation. However, a few measurable values indicate a better model:

- Less conflicting edges reduce the workload for manual review.
- The number of nodes for which a not-before or not-after limit could be inferred hints at the usefulness of the model.
- Different orderings of the witnesses can be compared, e.g. by using a rank correlation like Spearman's Rank Correlation based on the mean squared difference between the respective ranks of all nodes in two rankings.

Further work

Fig. 4 models witnesses as intervals. This could be improved by modelling *everything* as intervals and allowing all 13 possible relations in between. Allen (1983) developed a graph model for this, but that does not provide conflict resolution out of the box.

Vague absolute datings line "spring 1825" are normalized to standard crisp intervals. Research, e.g., by Holmen and Ore (2010) uses fuzzy sets to model fuzzy intervals, and Schockaert and De Cock (2008) define relations and inference on such fuzzy intervals. Whether this approach is tractable for our data still needs to be determined.

Open data

All source data and generated graphs as well as the software library developed to analyse and visualize them are publicly and freely available (sources and software at https://github.com/faustedition/, generated graphs at http://faustedition.net/macrogenesis)

References

Allen, James F. 1983. "Maintaining Knowledge About Temporal Intervals." Communication of Acm, 832-43.

Baharev, Ali, Hermann Schichl, and Arnold Neumaier. 2015. "An Exact Method for the Minimum Feedback Arc Set Problem," December, 34. http://www.mat.univie.ac.at/~neum/ms/minimum_feedback_arc_set.pdf.

Bohnenkamp, Anne. 1994. "... das Hauptgeschäft nicht außer Augen lassend": die Paralipomena zu Goethes "Faust". Frankfurt am Main: Insel Verl.

Brüning, Gerrit, and Oliver Hahn. 2017. "Goethes Helena-Dichtung in Ursprünglicher Gestalt. Zum Methodischen Verhältnis von Materialanalyse Und Textkritik." Editio 31 (1): 145-72. https://doi.org/10.1515/editio-2017-0009.

Fischer-Lamberg, Renate. 1955. "Untersuchungen Zur Chronologie von Faust II 2 Und 3." Diss. phil. (masch.), Berlin.

Goethe, Johann Wolfgang. 2019. Faust. Historisch-Kritische Edition. Edited by Anne Bohnenkamp, Silke Henke, and Fotis Jannidis. Version 1.2 RC. Frankfurt am Main ; Weimar ; Würzburg. http://faustedition.net/.

Hoenen, Armin, and Gerrit Brüning. 2019. "Zur Stemmatologie Neuerer Überlieferungen." DARIAH-DE Working Papers 29. urn:nbn:de:gbv:7-dariah-2019-1-3.

Holmen, Jon, and Christian-Emil Ore. 2010. "Deducing Event Chronology in a Cultural Heritage Documentation System - CAA Online Proceedings." http://proceedings.caaconference.org/paper/17_holmen_ore_caa2009/.

Kahn, A. B. 1962. "Topological Sorting of Large Networks." Commun. ACM 5 (11): 558-62. https://doi.org/10.1145/368996.369025.

Karp, Richard M. 1972. "Reducibility Among Combinatorial Problems." In Complexity of Computer Computations: Proceedings of a Symposium on the Complexity of Computer Computations, Held March 20-22, 1972, at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York, and Sponsored by the Office of Naval Research, Mathematics Program, IBM World Trade Corporation, and the IBM Research Mathematical Sciences Department, edited by Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, 85-103. The IBM Research Symposia Series. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-2001-2_9.

Schockaert, Steven, and Martine De Cock. 2008. "Temporal Reasoning About Fuzzy Intervals." Artificial Intelligence 172 (8): 1158-93. https://doi.org/10.1016/j.artint.2008.01.001.

Vitt, Thorsten, and Gerrit Brüning. 2019. "Determining and Visualizing Genesis: A Digital Edition of Goethe's Faust." In DH2019 Book of Abstracts. Utrecht. https://dev.clariah.nl/files/dh2019/boa/0924.html.

Wissenbach, Moritz, Dietmar Pravida, and Gregor Middell. 2012. "Reasoning About Genesis or the Mechanical Philologist." In Digital Humanities 2012: Conference Abstracts. Hamburg. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/reasoning-about-genesis-or-the-mechanical-philologist.1.html.

Taking a Visual Turn in DH

Lauren Tilton
ltilton@richmond.edu
University of Richmond, United States of America

Keywords: image analysis, computer vision, credit, institutions, big tent
Topics: Englis, hNorth America, 20th Century, Contemporary, image processing and analysis, meta-criticism (reflections on digital humanities and humanities computing), History, Media studies

The long paper will make an observation and offer a provocation. We are seeing and need a Visual Turn in DH. Scholars who identify as a part of DH have been pursuing what Münster and Terras have recently termed "visual digital humanities" for decades. Yet, this work has been less visible or ignored by the institutional structures that create, shape, and define "the big tent", what Jessica Marie Johnson terms "DHDH". In the sections that follow, I will focus on why and how we are seeing the Visual Turn in structures such as publications, conferences, grants, and awards. I argue that this turn has been in motion for decades but only recently occurring in DHDH. I'll finish with a provocation. We need a Visual Turn in order to center and credit exciting visual work as well as need to intervene in pressing questions about computational image analysis.

Algorithmically Mapping Historical Fascination in Late Imperial Chinese Literature

Paul August Vierthaler
pavierthaler@wm.edu
William & Mary, United States of America

Keywords: china, history, literature, bibliography, text mining
Topics: Asia, English, 15th-17th Century, 18th Century, bibliographic analysis, text mining and analysis, Asian studies, Book and print history

In late imperial China, historical narratives flourished across multiple genres of text, as authors were increasingly breaking free of the boundaries of standard historiographical forms and placing information in fictional and dramatic texts. The advent of large open-access digital corpora offers increasingly representative samples of imperial literary production, which I use to systematically study this phenomenon. In this paper, I algorithmically extract descriptions of historical events from a corpus consisting of historical, fictional, and dramatic texts dating from mid-Ming to high-Qing China (approximately 1500 to 1800). I track when imperial texts mention dates and people, extract when works are textuality intertwined, and

measure information density. I couple this with large-scale bibliographic analysis based on public domain library catalog records and contemporary bibliographies to map the historical focus of late imperial writing. I show that historical attention varies in genre specific ways and is dependent on myriad literary and social factors.

Indigenous Collaborations on Digital Public Memory Work in the Classroom

Amy J. Lueck
alueck@scu.edu
Santa Clara University, United States of America

Brian Beams
Santa Clara University, United States of America

Tanya Chiykowski-Rathke
Santa Clara University, United States of America

Matthew V. Kroot
Santa Clara University, United States of America

Keywords: Public Digital Humanities, Public Memory, Rhetoric, Indigenous Persistence, California Missions
Topics: English, North America, 18th Century, 19th Century, Contemporary, curricular and pedagogical development and analysis, rhetorical analysis, Anthropology, Literacy, composition, and creative writing

Communities across the United States and beyond are facing difficult decisions about how to represent the uglier sides of their histories. One place where this question has become prominent in recent years is on college campuses. In this paper, we argue that digital media embedded within collaborative pedagogical practices can be an effective way to intervene in public memory and historical representation. We specifically analyze our experiences working with local Native American communities in the development of curricular materials about Indigenous persistence on the campus of Santa Clara University during Spanish missionization. We show that (1) buy-in from the broader student body is a central ingredient in institutional change and (2) curricular interventions, using public-facing digital media projects, provide both a means of fronting Native voices and engaging a large student audience, in order to drive cultural transformations on campus.

Annotating Reader Absorption

Simone Rebora
simone.rebora@univr.it
University of Basel University of Verona

Piroska Lendvai
University of Basel

Moniek Kuijpers
University of Basel

Keywords: Reader absorption, Social reading, Inter-annotator agreement, Machine learning
Topics: Global, English, Contemporary, annotation structures, systems, and methods, text mining and analysis, Literary studies, Sociology

This paper presents our approach and interim results of the "Mining Goodreads" project, aimed at developing a computational approach to measure feelings of absorption in user-generated book reviews in English. Story world absorption is an experiential state driven by the reader's mental construction of a fictional world. Kuijpers (2014) proposed a theorization of story world absorption, distinguishing its four main dimensions: Attention (focused attention on the text, reducing awareness of the self and the passing of time), Transportation (the feeling of being transported to the story world), Emotional Engagement, and Mental Imagery. Kuijpers et al. (2014) developed and validated a corresponding questionnaire composed of 18 statements (Story World Absorption Scale, SWAS), which was already used in a series of studies on reader response (e.g., Hartung et al., 2017; Balint, Kuijpers & Doicaru, 2017; Kuzmicova et al., 2017).

Rebora et al. (2018) showed its application in DH research: manual analysis of 180 reviews published on Goodreads, one of the most popular "social reading" platforms (Cordón-García et al., 2013), indicated the utility of the SWAS in mining online book reviews for mentions of absorbing reading experiences. Multiple sentences of the Goodreads reviews show semantic and conceptual similarities with the 18 SWAS

statements. Our experiments with software for text reuse detection and textual entailment recognition indicated the need for improved quality and quantity of manual annotations to be used as training data.

Therefore, in our current project, funded by the Swiss National Science Foundation in the "Digital Lives" program (grant nr. 10DL15_183194), we pursue two aims based on labeled data obtained from five trained annotators: (1) data-driven validation of SWAS through confrontation with idiosyncratic, unstructured texts in which people describe their reading experiences; (2) development of an annotated corpus for training machine learning classifiers to detect different levels of absorption in millions of reviews hosted on social reading platforms.

So far, eight annotation rounds were completed on a total of 890 reviews, using the Brat (Stenetorp et al., 2012) and INCEpTION (Klie et al., 2018) annotation tools on fully flexible spans of text tokens. In the first round, we used six labels: SWAS_specific, for sentences that show direct similarity with the SWAS statements;SWAS_related, for sentences not included in the SWAS, but listed in a wider taxonomy of reading absorption (Bálint et al., 2016); mention_SWAS, for mentions of the SWAS concepts without reference to the actual reading experience of the user who wrote the review (i.e., "usually when I read a book, I like to be able to fully imagine what the world of the story looks like"); and candidate for unclassifiable, potentially relevant absorption phenomena. The first two labels also had a Present/Absent flag, for signaling sentences that explicitly confirm or negate absorption concepts. In the following rounds, these labels were extended to up to 145 tags, allowing to mark up more fine grained phenomena, like the four dimensions of story world absorption (Attention, Transportation, Emotional Engagement, Mental Imagery).

Inter-annotator agreement varies substantially depending on the labels. Figure 1 shows the evolution of Krippendorff's Alpha for the main tags in the eight rounds. As evident, there is a slight but steady improvement throughout the annotation process, that can be verified via the evolution of the "mean" and "all" scores: "mean" indicates the mean of the alpha scores for all of the tags (as it was not possible to calculate a single alpha score, because different tags could be assigned to the same sentences); "all" indicates the alpha score for a unique tag, obtained by checking if the sentence was annotated or not, independently from the assigned tag. In both cases, values move from fair (~0.2/0.4) to substantial agreement (~0.6/0.7). Among the high-level tags, SWAS_related_PRESENT reaches the highest values, while mention_SWAS_PRESENT scores the lowest, confirming the difficulty in recognizing absorption when no experiences of the I are mentioned.

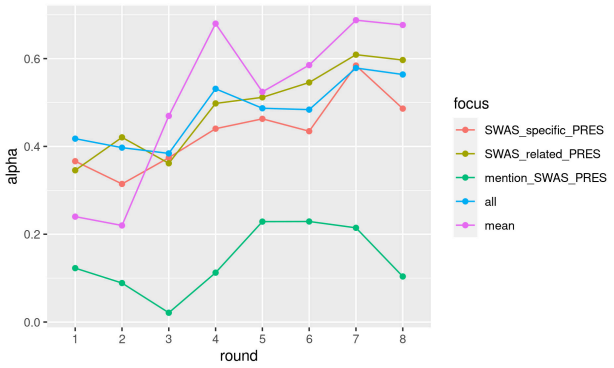


Figure 1. Inter-annotator agreement for the eight rounds of annotation. Alpha scores were calculated on a sentence basis (sentences split using spaCy)

Curation is currently in progress. However, first results confirm the already-observed trends. Mean agreement with the curator (mean Cohen's Kappa for the "all" tag) was 0.55 for the first round, while it reached 0.68 for the fourth.

The annotated segments offer the possibility to explore the different linguistic choices of reviewers. Figure 2 shows already some clear distinctions: verbs like "wait" and "want" and general references to the reading experience dominate the SWAS_related area. On the other hand, terms like "felt", "characters", and "story" indicate the stronger focus of the SWAS on personal involvement in the story world. These outcomes provide valuable information for validation and future development of the SWAS.

To confirm these findings, we used log-likelihood tests and compared the annotated texts with 100,000 randomly-selected reviews. Figure 3 confirms the dominance of emotional states in the SWAS_specific area, while SWAS_related terms like "wait", "next", "edge" and "seat" signal the frequent repetition of expressions of anticipation such as: "I cannot wait to read the next book in the series" and "I was on the edge of my seat".

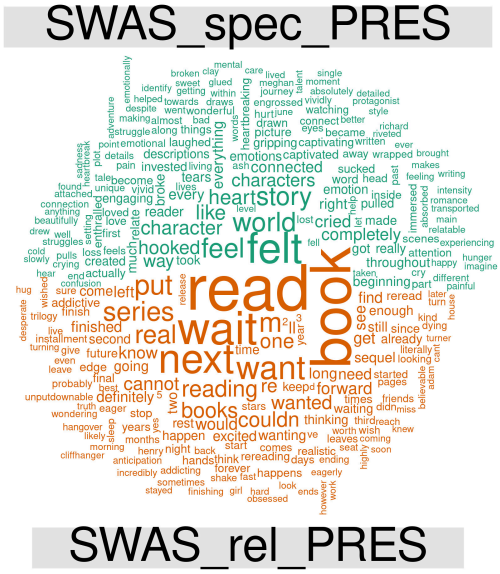


Figure 2. Comparative word cloud based on the most frequently annotated words (SWAS_specific_PRESENT vs. SWAS_related_PRESENT annotations)

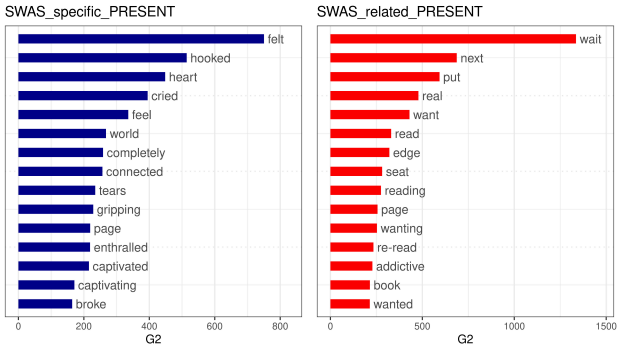


Figure 3. Overrepresented words in the annotated passages (SWAS_specific_PRESENT and SWAS_related_PRESENT annotations). Log likelihood calculated with the Quanteda R package

Focusing on the four SWAS dimensions, distinctions become even more fine grained. Figures 4 and 5 demonstrate for example how the phenomenon of Attention stimulates the widest reflections on the reading experience (e.g., on how "hooking" and "captivating" it can be). Transportation and Emotional Engagement partly share the keyword "felt": however, in the first case it means a more cognitively-aware "experiencing" of the fictional world, while in the second it relates to the deepest personal feelings (with a slight dominance of the negative over the positive). Mental Imagery shows at the same time the widest selection of words (cf. Figure 4) and its lowest distinctiveness (cf. Figure 5), confirming the lower presence and definition of the phenomenon in the Goodreads reviews.

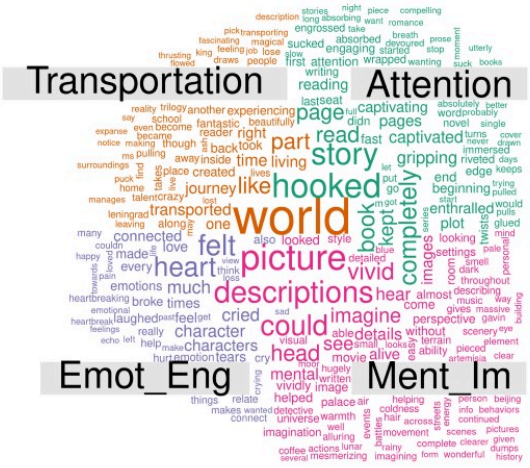


Figure 4. Comparative word cloud of most frequently annotated words (four SWAS dimensions)

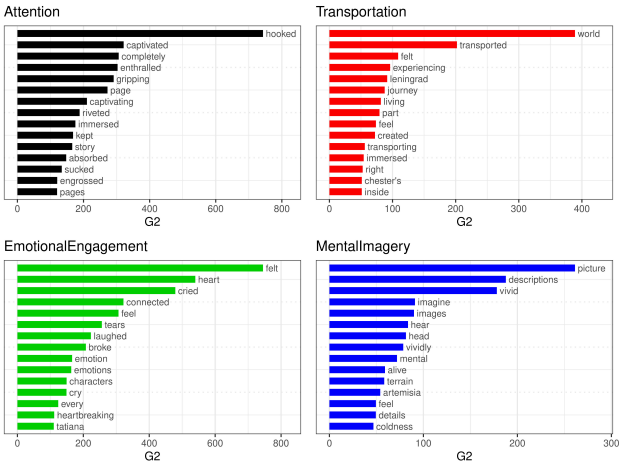


Figure 5. Overrepresented words in the annotated passages (four SWAS dimensions). Log likelihood calculated with the Quanteda R package

With the aim of extending the analysis on a much wider corpus, we used several state of the art machine learning approaches to train a binary classifier on the annotated reviews, cf. (Lendvai et al. 2020). When the current full dataset became available for training, a fine-tuned version of BERT (Devlin et al. 2018) reached 0.63 F-score on the target class, i.e., detecting absorption statements, and a linear regression model stacked on BERT predictions reached a mean average error of 0.08, cf. (Lendvai, Reichel, et al. 2020). These results strongly suggest that we can automate the annotation task in the future.

With reference to the conference theme, our research shows: (1) the utility of human agreement studies before the building of tools for DH; (2) the urge that computational literary studies move their attention towards the open and polyvocal treasure trove of information provided by social reading platforms (cf. Rebora et al., 2019).

References

Bálint, K., Hakemulder, F., Kuijpers, M. M., Doicaru, M. M. and Tan, E. S. (2016). Reconceptualizing foregrounding: Identifying response strategies to deviation in absorbing narratives. Scientific Study of Literature, 6(2): 176–207 doi:10.1075/ssol.6.2.02bal.

Bálint, K., Kuijpers, M. M., Doicaru, M. M., Hakemulder, F., & Tan, E. S. (2017). The effect of suspense structure on felt suspense and narrative absorption in literature and film. Narrative Absorption, 27, 177.

Cordón-García, J.-A., Alonso-Arévalo, J., Gómez-Díaz, R. and Linder, D. (2013). Social Reading. Oxford: Chandos.

Devlin, J., Chang M.W., Lee K., Toutanova K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, arXiv:1810.04805.

Hartung, F., Withers, P., Hagoort, P. and Willems, R. M. (2017). When Fiction Is Just as Real as Fact: No Differences in Reading Behavior between Stories Believed to be Based on True or Fictional Events. Frontiers in Psychology, 8 doi:10.3389/fpsyg.2017.01618. http://journal.frontiersin.org/article/10.3389/fpsyg.2017.01618/full (accessed 14 June 2020).

Klie, J.-C., Bugert, M., Boullosa, B., Castilho, R. E. de and Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Association for Computational Linguistics, pp. 5–9 http://tubiblio.ulb.tu-darmstadt.de/106270/.

Kuijpers, M. M. (2014). Absorbing stories. The effects of textual devices on absorption and evaluative responses. Unpublished PhD Thesis.

Kuijpers, M. M., Hakemulder, F., Tan, E. S. and Doicaru, M. M. (2014). Exploring absorbing reading experiences. Developing and validating a self-report scale to measure story world absorption. Scientific Study of Literature, 4(1): 89–122.

Kuzmičová, A., Mangen, A., Støle, H., & Begnum, A. C. (2017). Literature and readers' empathy: A qualitative text manipulation study. Language and Literature, 26(2), 137-152.

Lendvai, P., Darányi, S., Geng, C., Kuijpers, M., Lopez de Lacalle, O., Mensonides, J.-C., Rebora, S. and Reichel, U. (2020). Detection of Reading Absorption in User-Generated Book Reviews: Resources Creation and Evaluation. Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 4835–4841 https://www.aclweb.org/anthology/2020.lrec-1.595 (accessed 14 June 2020).

Lendvai, P., Reichel, U., Kuijpers, M. and Rebora, S. (2020). Ranking of Social Reading Reviews Based on Richness in Narrative Absorption. SwissText and Konvens 2020 5th SwissText & 16th KONVENS Joint Conference.

Rebora, S., Lendvai, P. and Kuijpers, M. (2018). Reader experience labeling automatized: Text similarity classification of user-generated book reviews. EADH2018. Galway: EADH https://eadh2018.exordo.com/programme/presentation/90 (accessed 14 June 2020).

Rebora, S., Boot, P., Pianzola, F., Gasser, B., Herrmann, J. B., Kraxenberger, M., Kuijpers, M., et al. (2019). Digital Humanities and Digital Social Reading. OSF Preprint doi:10.31219/osf.io/mf4nj. https://osf.io/mf4nj (accessed 14 June 2020).

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. and Tsujii, J. (2012). BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 102–07.

OuPoCo, the combinatorial poetry workbench (L'ouvroir de poésie combinatoire)

Thierry Poibeau
thierry.poibeau@ens.fr
ENS/PSL & CNRS, France

Mylène Maignant
ENS/PSL & CNRS, France

Frédérique Mélanie-Becquet
ENS/PSL & CNRS, France

Clément Plancq
ENS/PSL & CNRS, France

Matthieu Raffard
Atelier Raffard Roussel, France

Mathilde Roussel
Atelier Raffard Roussel, France

Mathilde Saurat
ENS/PSL & CNRS, France

Keywords: poetry; generation; rhyme analysis; digital poetry; user experience
Topics: Europe, English, 19th Century, electronic literature production and analysis, natural language processing, Literacy, composition, and creative writing, Literary studies

Oupoco (*L'ouvroir de poésie combinatoire*) is a project taking inspiration from Raymond Queneau's book *Cent mille mille milliards de poèmes*, published in 1961. Queneau's book is a collection of ten sonnets which verses can be freely recombined to form new poems. The book can be seen as composed of ten sheets, each separated into fourteen horizontal bands, each band carrying a verse on its front. The reader can choose, for each verse, one of the ten versions proposed by Queneau. The ten versions of each verse have the same scansion and rhyme, which ensures that each sonnet thus assembled is regular in shape [Queneau, 1961].

It would be tempting to develop a computer-based version of Queneau's work, but Queneau's book is still under copyright, and it is by definition limited to its ten original sonnets. To overcome this problem, we developed the Oupoco project, aiming at proposing a sonnet generator based on the recombination of a large collection of 19th century French sonnets. The challenge is thus more complex than the one proposed originally by Queneau since our sonnets do not have the same scansion and rhyme. From this point of view, even if the project is intended to generate new sonnets, it is largely based on the development of analysis tools able to identify the scansion, the rhyme and the structure of the original sonnets. It is thus very different from the numerous projects dedicated to the pure generation of poetry, being with symbolic [Gervás, P., 2013] or neural methods [Ghazvininejad et al., 2017] [Van De Cruys, 2019] (among many others).

Oupoco is currently based on a collection of 788 sonnets from 16 authors from the 19th century, and this database is regularly expanding. Each sonnet is encoded in a XML format along with related metadata, and a TEI version of the database is available. The project requires to get access to a formal representation of rhymes [Beaudouin, 2002]. In order to do this, the first step is to get a phonetic transcription of the last word of each verse, but this is not enough: for example, aimé and aimée have the same phonetic transcription, but do not rhyme, according to French rhyming rules (feminine and masculine do not rhyme according to the classical rules of French poetry); there are also cases where the phonetic transcription diverges but words actually rhyme (for example with sounds like [e] and [e]). A series of rules had thus to be defined to get a proper analysis of rhyme from the phonetic transcription of the last word of each verse. The generator uses this analysis to produce random sonnets, with different possible structures, respecting the rules of French versification (the code and the resources used, especially the sonnet database, are open source and freely available for research, see: https://github.com/clement-plancq/oupoco-api).

A series of "side products" have been produced from the project, including:

- a website (https://oupoco.org/):

Figure 1: Web page introducing the Oupoco website

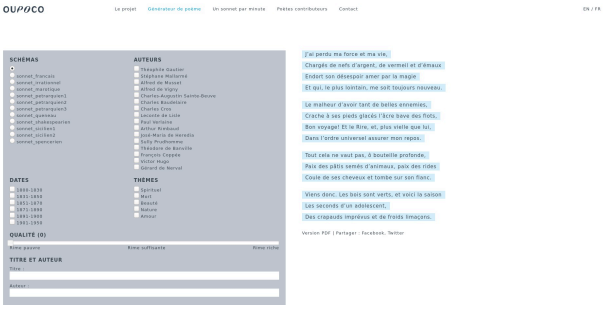


Figure 2: A generated poem, along with the constraint panel (on the left), allowing the user to select the authors and the structure of the poem he wants to generate. Placing the mouse on a given verse gives access to the exact reference of that verse.

- a bot posting on Twitter (https://twitter.com/ oupoco_bot):

Figure 3: The Oupoco bot on Twitter, generating a quatrain every 6 hours.

- and a "poetry box" (*la boîte à poésie*), a portable version of the original idea that can be demonstrated in public events (based on Raspberry Pi components). Through these devices our goal is to reach a wider audience and engage people to reconnect with poetry

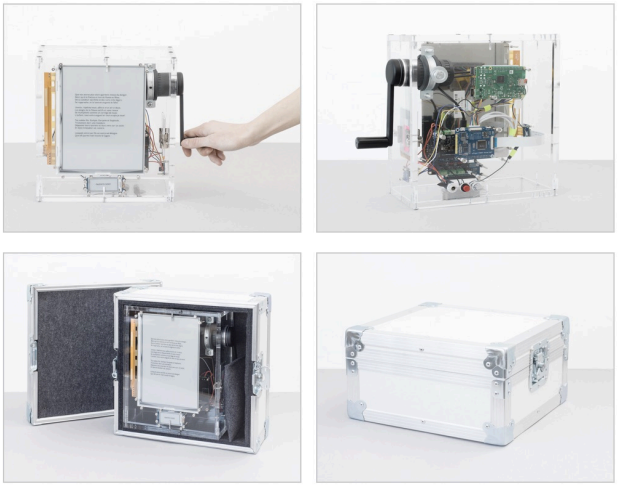


Figure 4: « La Boîte à poésie », a portable version of the project, conceived Atelier Raffard-Roussel (Paris). This portable device allows one to demonstrate the system in various venues; the electric power required to generate a poem can be produced manually, thanks to the crank on the side.

The main interest of the project is to present French poetry through a new and original setting. With our system, poetry is not any more just a literary genre [Derrida and Ronnel, 1980], but a dynamic object that can be manipulated and experienced. For lots of people, poetry is seen at best as something related to school years, at worse as something boring and uninteresting from the past. Our new setting, in itself, makes it possible to show that playing with poetry can be fun. Our setting puts in perspective the notion of text coherence [Reinhart, 1980], since the result of the generator can be more or less satisfactory from a semantic point of view.

This has two consequences. The first one is related to interpretation: because the machine produces structurally impeccable sonnets, the experiencer is unconsciously encouraged to find coherence in them, simply because we are used to coherence in our everyday life and because incoherence is bewildering. The second consequence is a frequent need for the experiencer to go back to the original poem, to see where from a given verse originates (tooltips always allows the experiencer to go back to the original sonnet where a verse has been extracted). The project is thus not just a sacrilege game over venerated texts, but a way to make people experience and rediscover poetry.

References

Beaudouin, V. (2002). Mètre et rythme du vers classique. Corneille et Racine. Champion (coll. Lettres numériques) : Paris.

Derrida, J. and Ronell, A. (1980)."On Narrative: The Law of Genre". In Critical Inquiry. The University of Chicago Press. 7 (1): 55–81. doi:10.1086/448088.

Ghazvininejad, M., Shi, X., Priyadarshi, J. and Knight K. (2017). "Hafez: an Interactive Poetry Generation System". Proceedings of ACL 2017, System Demonstrations, Vancouver, Canada. pp. 43–48.

Gervás, P. (2013). "Computational Modelling of Poetry Generation", in Artificial Intelligence and Poetry Symposium, AISB Convention 2013, University of Exeter, United Kingdom.

Queneau, R. (1961). Cent mille milliards de poèmes. Gallimard : Paris.

Reinhart T. (1980). "Conditions for Text Coherence". In Poetics Today, Vol. 1, No. 4 (Narratology II: The Fictional Text and the Reader), Duke University Press, pp. 161-180. doi: 10.2307/1771893.

Van de Cruys, T. (2019). "La génération automatique de poésie en français". Proceedings of TALN 2019, Toulouse, France. pp. 113–126.

Acknowledgement

This work has received support of Translitteræ (Ecole universitaire de recherche, program "Investissements d'avenir" ANR-10-IDEX-0001-02 PSL* and ANR-17-EURE-0025) and the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

An Ontology for Critical Editions of Variant Text

Tara Lee Andrews

tara.andrews@univie.ac.at
University of Vienna, Austria

Keywords: critical edition, ontology, modelling
Topics: Asia, Global, Europe, English, BCE-4th Century, 5th-14th Century, 15th-17th Century, data modeling, scholarly editing and editions development, analysis, and methods, Informatics, Philology

This contribution proposes an ontology that arises from work on the Stemmarest text tradition repository, carried out as part of the SNSF project "The Chronicle of Matthew of Edessa Online". We describe the definitions, relationships, and assumptions that have been made in the process of producing a functional piece of software whose primary object is to collect and manipulate data in order to produce a critical edition of variant text. Our purpose is twofold: first, to offer a resource (or at least a starting point) for those who would develop software for text critical edition; second, as a practical experiment, to explore where our data model tests the limits of ontological reasoning.

Sustainably modeling a 3D virtual tour: Can we avoid the "black box"?

Elisa Eileen Beshero-Bondar

ebb8@pitt.edu
University of Pittsburgh at Greensburg, United States of America

Pilar Maria Herr

pmh3@pitt.edu
University of Pittsburgh at Greensburg, United States of America

Amber A. McAlister

University of Pittsburgh at Greensburg, United States of America

Gregory H. Bondar

ghb2@pitt.edu
University of Pittsburgh at Greensburg, United States of America

Joseph Bleehash

University of Pittsburgh at Greensburg, United States of America

Keywords: 3D, virtual modeling, GIS, photogrammetry, sustainable design
Topics: EnglishNorth America19th Century20th Century, Contemporary, spatial & spatio-temporal analysis, modeling and visualization, virtual and augmented reality creation, systems, and analysis, Geography and geo-humanities, History

The authors are evaluating 3D game engine software to integrate GIS topographic survey data with photogrammetry as they build a virtual tour of a heritage site. The virtual tour will describe building exteriors and interiors, plus historical artifacts within the buildings that visitors should be able to examine from all angles on screen. The interface must be lightweight to run

smoothly on mobile devices and must efficiently filter out information to guide the visitor to focal points. We seek to avoid the "black box" problem of locking the data in interface software, and rather, our data must be (con)served in openly accessible and readily queryable formats like CityGML, CityJSON, or simple CSV files. The authors will discuss their decisions for sustaining and sharing the project data, and their evaluation of 3D game engine software for its capacities to translate spatial data and to support annotation and user interactivity.

Genocidio de los pueblos originarios y modos de distribución de las tierras fiscales en la Patagonia Austral: la prensa local y los regímenes de propiedad agraria (1929-1962).

gustavo urbano navarro
gusunavarro@gmail.com
Unpa-Uasj, Argentine Republic

Keywords: Indigenismo, prensa, visualización de la información, extracción de datos
Topics: South America, Spanish, 19th Century, 20th Century, text mining and analysis, Humanities computing, Media studies

La historia de Patagonia está atravesada por procesos de dominación, genocidio de sus pueblos originarios y movimientos migratorios. La prensa desempeñó allí un papel fundamental ya que sus élites crearon medios de prensa dirigidos a defender sus intereses. En ellos, aparecen de forma recurrente distintos regímenes de propiedad agraria.

El proyecto identifica en la prensa local, los modos de distribución de la tierra en la región en el período de su poblamiento definitivo. Realiza minería de datos en los periódicos para graficar el cruce de ambas dimensiones. La selección del material documental respondió al criterio de "nivel de difusión" de las publicaciones: Periódico "La Verdad y El Sonido" (1914-1939), Periódico "El Orden" (1912-1960).

La metodología se basó en el análisis computacional para ilustrar la conexión entre ambas dimensiones. Se utilizó "Pandas", una librería de python destinada al análisis que proporciona estructuras de datos flexibles. Luego, se utilizó "Matplotlib" para su visualización.

Training Algorithms to Read Complex Collections: Handwriting Classification for Improved HTR Models

Carrie Pirmann
carrie.pirmann@bucknell.edu
Bucknell University, United States of America

Bhagawat Acharya
bhagawat.acharya@bucknell.edu
Bucknell University, United States of America

Brian King
Bucknell University, United States of America

Katherine Mary Faull
Bucknell University, United States of America

Keywords: machine learning, paleography, manuscript studies, collaboration, undergraduate research
Topics: Europe, English, North America, 18th Century, 19th Century, artificial intelligence and machine learning, Computer science, Humanities computing

This paper will present a new handwriting grouping algorithm that has been developed to decrease the Character Error Rate (CER) for a collection of manuscript documents written in various hands and in multiple languages.

The Moravian Lives project (moravianlives.org), an international, collaborative DH project housed at Bucknell University, takes as its starting point a vast collection of archival materials held by the Moravian Church in the United Kingdom, Germany, and the United States. The materials include tens of thousands of handwritten ego-documents, written in a variety of handwriting styles. As the documents are held in

archives in a variety of international locations, one of the goals of the Moravian Lives Project is to digitize and transcribe these memoirs, to make them accessible to a broader audience.

Initially, transcriptions for the memoirs were crowdsourced. However, crowdsourcing is replete with problems including varying accuracy of transcriptions, length of time needed to produce transcriptions, and a dearth of individuals who can read the handwriting styles of the documents, particularly those written in old German script.

To facilitate the transcription process, in early 2019 the Moravian Lives team began using Transkribus (transkribus.eu). The platform allows for creation of custom handwritten text recognition (HTR) models, which are based on previously transcribed memoirs and used to machine transcribe new documents (Muehlberger et al., 2019). With adequate training data (i.e., several hundred pages or 50,000+ words), models with a CER of five percent or less can be developed, which is sufficient for expediting archival work. Extant projects which have so far achieved this success rate may be based on multiple hands, drawing on significant data from each hand. For example, the University of Greifswald has trained successful models with a 5% CER on a corpus of 250,000 words written in three different hands. Similarly, the Bentham Project trained a highly accurate English-language model on 50,000 words written in a small number of hands (Muehlberger et al., 2019).

The numerous and varying handwriting styles found in the Moravian memoirs present multi-facted challenges to creating highly accurate models. We do not know how many scribes there were, or in most cases their identities, and we are continually coming across new handwriting styles. Memoir documents are between two to 50 pages in length; most documents we are working with are 10 pages or fewer, meaning there is not a lot of data per document. While we have had some success creating models via human identification of similarities in handwriting, we believe that automated scribe identification and/or automated grouping of handwriting by similarities in style could result in much more accurate models. To address this problem, an undergraduate computer science major and professor of computer science joined the Moravian Lives team and are experimenting with deep learning to author a grouping model, designed to group or sort memoirs by handwriting styles. These groupings should enable the creation of more accurate models in Transkribus, as well as more accurate transcription outputs.

Bibliography

Fischer, Andreas, Micheal Baechler, Angelika Garz, Marcus Liwicki, and Rolf Ingold. "A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents." In 2014 11th IAPR International Workshop on Document Analysis Systems. IEEE, 2014.

Kahle, Philip, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. "Transkribus-A Service Platform for Transcription, Recognition and Retrieval of Historical Documents." In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2017.

Mohammed, Hussein, Volker Märgner, and H. Siegfried Stiehl. "Writer Identification for Historical Manuscripts: Analysis and Optimisation of a Classifier as an Easy-to-Use Tool for Scholars from the Humanities." In 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2018.

Moravian Lives. Bucknell University and the University of Gothenburg, 2019, http://moravianlives.org/ Accessed 20 October 2019.

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto et al. "Transforming Scholarship in the Archives Through Handwritten Text Recognition." Journal of Documentation, vol. 75 no. 5, 2019, pp. 954-976.

Senatsprotokolle der Universität Greifswald. University of Greifswald, 2019, http://www.digitale-bibliothek-mv.de/viewer/toc/PPNUAG_0_1/1/LOG_0000/ Accessed 20 October 2019.

Transcribe Bentham. University College, London, 2019, http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham Accessed 20 October 2019.

Transkribus, University of Innsbruck, 2019, https://transkribus.eu/ Accessed 20 October 2019.

Ventresque, Vincent, Arianna Sforzini, and Marie-Laure Massot. "Transcribing Foucault's handwriting with Transkribus." Journal of Data Mining & Digital Humanities, 2019. https://jdmdh.episciences.org/5218/pdf

Queering the Tape Recorder: Transforming Surveillance Technologies through bill bissett's Queer Poetic Voice

Mathieu Aubin
mathieu.aubin25@gmail.com
Concordia University, Canada

Keywords: literary audio; queer; surveillance; technology
Topics: English, North America, 20th Century, music and sound digitization, encoding, and analysis, Gender and sexuality studies, Literary studies

This presentation reports on my preliminary analyses of a digitized collection of literary audio recordings featuring poetry readings by Canadian sound poet bill bissett. As Canadian literary communities adopted audio recording technologies as part of their cultural practices during the 1960s (Davidson 203), bissett developed a unique relationship to the tape recorder. At the time, the RCMP used this technology to listen to queer people's conversations, including those of bissett, document their activities, and regulate their sexuality in order to ensure the nation's heterosexual status quo (Kinsman and Gentile 29). In the poet's tape recordings ranging from the 1960s-1980s, we hear bissett read poetry theorizing this queer surveillance, documenting his lived experiences as a gay man, and fighting for sexual liberation. We also hear sonic traces of the location and social dynamics of the recorded event and the technology used to capture this moment. Until recently, when SpokenWeb began digitizing and making them available to the wider public, these recordings were kept in private collections and had limited circulation. Following the recordings' recent shift towards digital public circulation, this presentation considers how listening to bissett's queer tape recordings in SpokenWeb's digital archive amplifies his voice, forges queer ways of listening to literary audio, and fosters new public dialogue about Canada's gay history.

SpokenWeb is a multi-institutional research network that digitizes and connects tape and reel-to-reel recordings through its online repository, while creating an online sound archive open to the general public. Audio objects such as these recordings, Canadian sound studies scholars have argued, are significant literary artifacts for investigating previously overlooked Canadian literary histories (Camlot and Marshall; Fong and Shearer). I suggest that SpokenWeb's digital archive, featuring gay recordings, functions as a site of queer social resistance. As feminist and queer scholars argue, the archive can be a site for queer ethical interventions, allowing us to better understand the concerns and ways of being of queer people in the past and the present (Cvetkovich; Eichhorn). In developing a feminist close listening practice and applying it to digitized literary audio recordings featuring women within SpokenWeb's collections, Deanna Fong and Karis Shearer have amplified women's contributions to Canadian literature and determined how women's labour has been erased from previous literary histories. Whereas Fong and Shearer's work focuses on listening for women's contributions within digitized recordings, this presentation shows how listening queerly to SpokenWeb's gay recordings activates its queer activist potential.

Digital humanities scholars have deployed a variety of digital tools to analyze audio by marginalized communities (e.g., pitch analysis [MacArthur]; sound visualization [Hammond and Dick]). However, critical race and sound studies scholar Nina Sun Eidsheim critiques the use of machine learning for engaging with people of colour's voices, arguing that doing so reproduces racist assumptions of voice (115-117). Along the same critical line, I contend that the use of conventional digital sound analysis tools to analyze sound in bissett's recordings would reproduce the essentialist stereotypes about gendered voice akin to those of the RCMP's "fruit machine." As an alternative approach, I suggest a qualitative listening methodology in which I listen for queer sonic resonances in bissett's digitized recordings as well as new digital born oral literary history recordings that I produce with the poet. Here, the term "queer" reflects an identity marker for people who do not identify with heteronormative assumptions of gender and sexuality and the term "sonic" means what is audible to the human ear. Informed by queer oral historian practices (Boyd and Ramírez), listening for queer sonic resonances focuses on how bissett interprets and articulates his life through his poetry, uses queerly-coded language based on the event's context, builds relationships with audiences through discussions of culture, intimacy, friendship, politics, and sexuality, and creates moments of dissonance through silences. To return to the poet's subjectivity and let the poet speak in response to the archive, I have also conducted and recorded oral history interviews with bissett in which we discussed the earlier recordings' personal and historical contexts and imagined how these new oral history recordings could provide new historical information by being placed in SpokenWeb's digital archive. Thus, my research so far has activated the queer activist potential of SpokenWeb's archive by imagining these recordings not as data to mine but rather as bodies of evidence of queer lived experiences that resist the heteronormative status quo.

Through my presentation's analysis of the recordings, I suggest that engaging queerly with

the digital literary audio archive and producing new queer audio reveals contextual information not available through digital analysis tools. Specifically, I compare different performances of bissett's poem "th wundrfulness uv th mountees our secret police," which documents his experience of being surveilled and tape recorded by the RCMP. This comparison shows how bissett accentuated different parts of the poem depending on the event, how audiences responded to his work, and how he engaged with tape-recording technology. I also demonstrate how conducting oral history interviews with bissett over the past two years and engaging with the recordings has enlivened them and provided important contextual information about those historical moments. For instance, in previous oral histories interviews when we discussed recordings of "th wundrfulness uv th mountees our secret police," bissett spoke about his experience of being ostracized by Vancouver's conservative publics and the Canadian government. These digital born recordings form oral testimonies that produce intergenerational dialogue about Canadian queer history. As this paper ultimately contends, listening queerly to the digitized earlier recordings and conducting oral history interviews with bissett activate the queer activist potential of SpokenWeb's digital sound archive, thereby empowering bissett in this current moment.

Works Cited

Boyd, Nan Alamilla, and Horacio N. Roque Ramírez, editors. Bodies of Evidence: The

Practice of Queer Oral History. Oxford UP, 2012.

Camlot, Jason, and Christine Marshall, editors. "The Poetry Series." Amodern, vol. 4, 2015.

Cvetkovich, Ann. Archive of Feelings. Duke UP, 2003.

Davidson, Michael. Ghostlier Demarcations: Modern Poetry and the Material Word. U of

California P, 1997.

Eichhorn, Kate. The Archival Turn in Feminism: Outrage in Order. Temple UP, 2013.

Fong, Deanna, and Karis Shearer. "Gender, Affective Labour, and Community-Building

Through Literary Audio Artifacts." No More Potlucks, 2018,

http://nomorepotlucks.org/site/genderaffective-labour-and-community-building-through-

literary-audio-artifacts-deanna-fong-andkaris-shearer/. Accessed 1 June 2020.

Hammond, Adam, and Jonathan Dick. "'th mold thats branded on m soul': A Computational

Approach to Racialized Voice in Jean Toomer's 'Kabnis.'" Resonant Practices in

Communities of Sound, The SpokenWeb Symposium, 31 May 2019, Harbour Centre,

Vancouver, BC. Conference Presentation.

Kinsman, Gary, and Patrizia Gentile. The Canadian War on Queers. UBC P, 2010.

MacArthur, Marit. "Introducing Simple Open-Source Tools for Performative Speech Analysis:

Gentle and Drift." Jacket 2, 6 Jun. 2016, https://jacket2.org/commentary/introducing-

simple-open-source-tools-performative-speech-analysis-gentle-and-drift. Accessed 12

Jan. 2020.

Finding Identical Ukiyo-e Prints across Databases in Japanese, English and Dutch

Yuting Song

songyt@fc.ritsumei.ac.jp
Research Organization of Science and Technology, Ritsumeikan University, Japan

Biligsaikhan Batjargal

Kinugasa Research Organization, Ritsumeikan University, Japan

Akira Maeda

College of Information Science and Engineering, Ritsumeikan University, Japan

Keywords: Ukiyo-e, cross-language record linkage, bilingual word embeddings
Topics: Global, English, 18th Century, natural language processing, Computer science, Library & information science

Many cultural memory institutions provide online access to their collections. Ukiyo-e, Japanese traditional woodblock print, is known as a popular art of the Edo period (1603-1868). There are many copies printed from the same ukiyo-e woodblocks, and they have been digitized and exhibited on the Internet by many museums around the world with descriptive metadata in various languages. This means that identical ukiyo-e prints can exist in different digital collections, along with metadata in different languages. This proposal aims at finding these identical ukiyo-e prints by comparing their metadata across languages.

We measure the metadata similarities across languages by utilizing bilingual word embeddings, that allow measuring word similarities across languages. Our method can be used to link or integrate digital collections in different languages, which could help people to explore ukiyo-e prints regardless of their metadata's languages.

Pulses: Decentralized Smooth Semantization of Atomic Historical Knowledge

Frédéric Kaplan
frederic.kaplan@epfl.ch
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Raphael Barman
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Mathieu Clavel
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Albane Descombres
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Hakim Invernizzi
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Marion Kramer
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Laurene Puerto
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Santiago Saint-supery
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Cedric Viaccoz
Ecole Polytechnique Fédérale de Lausanne, Switzerland

Keywords: Twitter, Wikipedia, Semantic web, Digital History
Topics: Europe, English, Contemporary, data modeling, public humanities collaborations and methods, Computer science, History

This paper presents the prototype of a platform aimed at curating simple open bricks for historical knowledge, shaped as short texts. These historical bricks, named "Pulses," are meant to be based on no prefixed ontology, but collectively negotiated conventions and reformatting by bots. The Pulses curation platform that can be interpreted as hybridisation between some characteristics of Twitter and Wikipedia, organises a collective process of "smooth semantization" in which pulses are progressively rewritten to become more machine readable. The article reports on the principle and implementation of this system conducted in the last two years.

Encoding Working Lives: Linking Labor, Office, and Religion in 18th-century Manuscript Collections

Diane K. Jakacki
dkj004@bucknell.edu
Bucknell University, United States of America

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Justin Schaumberger
jes083@bucknell.edu
Bucknell University, United States of America

Keywords: moravian, 18th century, industrial revolution, Yorkshire
Topics: Europe, English, 18th Century, data modeling, text encoding and markup language creation, deployment, and analysis, Cultural studies, History

This paper will detail the development of an ontology that reflects the complex relationship between concepts of labor and office in the 18th century Moravian ego-documents of Fulneck, Yorkshire. Undertaking this type of rich and subject-driven data mining permits the exploration of the relationship between forms of labor in Moravian congregations and their relationship to a broader understanding of work in the 18th century. It also provides a platform for further investigation of contemporary religious societies.

Historically, the membership of the Moravian Church in the United Kingdom, like the Methodist Church of the Wesleys, has been seen as consisting primarily of men and women from the working class whose artisan and laborer skills were fundamentally redeployed by the advent of large-scale industrial production in the North of England. Thanks to the Moravian custom of writing a memoir these working people left autobiographical documents that include details of their working lives. For the most part unstudied, these English-language memoir collections provide historians with a treasure-trove of new archival material, written by the working-class members of the church.

Recent scholarship on the intersection of gender, industry and religion (See Hammond and Forsaith, 2011) has begun to explore how these concepts were transformed by the early Industrial Revolution and provide a welcome revision to the dominant historiographical narrative of the British working class. (As inscribed by E.P. Thompson, Making of the English Working Class, 1963.) This corpus of unpublished and mostly uncatalogued material allows us to examine the sentiments directed towards labor and office respectively.

Since 2014, the Moravian Lives project (moravianlives.org) has undertaken the work of transcribing, encoding, and data-mining the memoirs written by thousands of Moravians in Europe, North America, and the Caribbean. Recently the project team has turned its attention to the English congregations, with a focus on revealing prosopographical information within of the memoirs.

Using the Fulneck memoirs as our corpus, we have begun the process of harvesting this prosopographical data through semantic text encoding detailed personography, minting unique identifiers and capturing data about key life and religious events. In the process of so doing, we have realized that we were also producing unique ontology of Moravian offices and occupations undertaken by these Moravians, within their community and within broader regional context. (See Eide, and Ciotti and Tomasi for an outline of challenges of using TEI XML for this kind of intensive semantic work.and data modeling.)

By July 2020 the Moravian Lives project team will be able to report on the status of the prosopography and ontology. More importantly, we will be able to present initial analysis of the connections this work enables us to reveal - between male and female members of the "economies", the shifts we detect in types of labour from the 1740s through the 1790s, and how this methodological approach might be used as a model for other research projects focusing on concurrent evangelist movements, and those concentrating more broadly on related themes and trends to do with the Industrial Revolution.

Bibliography:

Bingenheimer, Marcus, Jen-Jou Hung, and Simon Wiles. "Social network visualization from TEI data." Literary and Linguistic Computing. (2011): 271-78.

Ciotti, Fabio, and Francesca Tomasi. "Formal Ontologies, Linked Data, and TEI Semantics." Journal of the Text Encoding Initiative, no. Issue 9, Sept. 2016. journals.openedition.org, doi:10.4000/jtei.1480.

Eide, Øyvind. "Ontologies, Data Modeling, and TEI." Journal of the Text Encoding Initiative. no. Issue 8, Dec. 2014. journals.openedition.org, doi:10.4000/jtei.1191

Freund, Alexander. Beyond the Nation : Immigrants' Local Lives in Transnational Cultures. University of Toronto Press, 2012.

Hammond, Geordan, and Peter S Forsaith. Religion, Gender, and Industry: Exploring Church and Methodism in a Local Setting. James Clarke Company, Limited, 2012.

Pasin, Michele and John Bradley. "Factoid-based prosopography and computer ontologies: towards an integrated approach", Literary and Linguistic Computing, Volume 30, Issue 1, April 2015, Pages 86-97.

Stead, Geoffrey. The Moravian Settlement at Fulneck, 1742-1790. Thoresby Society, 1999.

Thompson, E. P. The Making of the English Working Class. Pantheon Books, 1964.

WoPoss - A World of Possibilities. An open access annotated corpus to study the diachrony of modality in the Latin language

Francesca Dell Oro
francesca.delloro@unil.ch
Université de Lausanne, Switzerland

Keywords: diachronic semantics, modality, fine-grained semantic annotation, corpus linguistics
Topics: GlobalEnglishBCE-4th Century, Contemporary, annotation structures, systems, and methods, semantic analysis, Linguistics, Philology

Until now the study of lexical modality in the Latin language has focused almost exclusively on the synchronic level and mainly on verbs. The investigated corpus has also been limited to a small group of classical authors. Later sources or documentary sources, such as inscriptions, have been neglected. Moreover, no fine-grained annotation of modal types and subtypes has yet been attempted.The SNSF-funded project WoPoss - A World of Possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language (agreement n° 176778) aims to fill these gaps and to make available to the scientific community an annotated corpus which will be useful for research on modality or modality-related issues. As annotation progresses, data will be made available during the project lifespan (February 2019-January 2023). In this paper I want to present the main innovative features of the WoPoss corpus and of the WoPoss annotation.

Reconstructing the gendered division of labor in the French textile trades. Distant reading of primary qualitative sources with NLP tools (18th century-beginning of the 20th century)

Alix Chagué
alix.chague@inria.fr
Inria, France

Marie Puren
marie.puren@cnrs.fr
Laboratoire de recherche historique Rhone-Alpes / CNRS, France

Manuela Martini
Laboratoire de recherche historique Rhone-Alpes / Université Lumière Lyon 2, France

Eric de La Clergerie
Inria, France

Pierre Vernus
Laboratoire de recherche historique Rhone-Alpes / Université Lumière Lyon 2, France

Lionel Tadjou
Inria, France

Keywords: Textile Industry, Women Labor, Handwritten Text Recognition, Information extraction, French
Topics: Europe, English, 18th Century, 19th Century, 20th Century, natural language processing, text mining and analysis, Gender and sexuality studies, History

The study of the division of labour is undoubtedly one of the major topics in the history of labour and industrialization. In general, it has been addressed through the quantitative analysis of occupational

categories in censuses, lists of professions or accounting records. This paper proposes an alternative approach, following new studies linking social history, digital humanities and tools to process natural language, and using a semi-massive corpus of primary qualitative sources to deal with this issue. We present how historians, computer scientists and NLP experts work together on this topic within the frame of the TIME-US research project, focusing, on the one side, on the methodology set to implement distant reading and, on the other side, on the resulting tools developped to support historians' research hypotheses.

Crowdsourced Text Transcription on the Zooniverse Platform: Models for Design, Development and Evaluation

Samantha Nicole Blickhan
samantha@zooniverse.org
The Adler Planetarium, United States of America

Victoria Anne Van Hyning
vvanhyning@loc.gov
Library of Congress

Keywords: crowdsourcing, tool development, text analysis, methods, comparative analysis
Topics: Global, Europe, English, North America, 15th-17th Century, 19th Century, Contemporary, crowdsourcing, text mining and analysis, Design studies, History

This paper will discuss the steps taken by the team behind the Zooniverse.org crowdsourcing platform to address the relationship between task design and project outcomes, including a range of approaches to text-based data collection, volunteer engagement, and evaluation of methods alongside results. By focusing on two 'generations' worth of online, crowdsourced text transcription projects, the presenters will show how the Zooniverse team has used the successes--and failures--of previous projects as the basis for newly-developed methods of data collection. The presentation will focus on two projects launched in 2015 and their results, and will show how the methods used by those projects inspired the creation of an A/B experiment that ran in 2018. The authors will discuss the experiment results, and how the results affected the tool development for the platform that is currently underway.

Teaching at the Intersection of Digital Humanities and Visualization

Stefan Jänicke
stjaenicke@imada.sdu.dk
University of Southern Denmark, Odense, Denmark

Keywords: teaching, visualization
Topics: Global, English, Contemporary, curricular and pedagogical development and analysis, spatial & spatio-temporal analysis, modeling and visualization, Computer science

The increasing value of information visualization techniques to support investigating quantitative research questions in humanities applications is well-documented (Jänicke et al., 2017; Windhager et al., 2019). At the same time, it is important to make (digital) humanities scholars literate in dealing with visualizations to ensure that accurate conclusions can be drawn. In order to make younger generations ready for interdisciplinary work in a digital humanities context, I taught a module that attracted both computer science and humanities students. This article reflects on the most important aspects in teaching the course throughout three years.

Cohort of students

Teaching needs to be flexibly organized in dependency on the backgrounds of students joining a course. A course might be offered exclusively for either humanities or computer science students, or it can attract students with diverse study subjects. I faced different constellations, and emphasized different aspects according to the demands of the cohort. Whereas humanities students profit from more intense discussions on computational thinking and data modeling, it is especially helpful for computer science students to get to know typical research interests and traditional workflows of humanities scholars. The major

focus should be to ensure that students, independent of their backgrounds, learn to "speak the same language" using the same terminology.

Teaching visualization theory

To serve students with an easy-to-digest overview of visualization design, I recommend Tamara Munzner's book "Visualization Analysis and Design" (Munzner, 2014). It provides an introduction to data and task abstraction that are necessary to comprehend and to develop new or adapt existing visualizations. It further discusses how data features can be appropriately mapped to visual features, and how users can interact with visualizations. Also, emphasis should be devoted to Shneiderman's Information Seeking Mantra "Overview first, zoom and filter, then details-on-demand" (Shneiderman, 1996) as it encapsulates the general idea of quantitative data analysis without losing the materials a visualization is composed of. This rather theoretical frame should be accomplished with discussing visualization techniques that are of particular importance for digital humanities research: geographical maps, timelines, tag clouds, heat maps and graphs.

Ready-to-use-based vs. development-driven visualizations

Next to theoretical contents, the course should have a strong focus on practical work. It is known that different approaches to make use of visualization for knowledge discovery in digital humanities applications exist (Jänicke, 2016). The first approach is to apply ready-to-use tools like Voyant (Sinclair and Rockwell, 2020) for quantitative visual text analysis or Gephi (Bastian et al., 2009) for graph visualization. While the advantage of using such frameworks is generating arguable visual output in a short amount of time, research interests might deviate from what the tool can provide. Further, scholars need to learn how to use potentially complex tools and how to interpret upcoming results. In contrast to applying existing tools to generate visualizations, (digital) humanities and visualization scholars might also engage with each other aiming to generate a new visual vocabulary to expedite knowledge discovery, thereby facing the problems of interdisciplinary collaborations. While the first approach should be carried out simultaneously to the theoretical sessions, I recommend to conduct interdisciplinary student projects in the second part of the course.

Supervisory roles during project work

Training to apply the learned visualization-related terminology should be the main focus of a project, while the visualization result itself plays a secondary role. Especially in a setting with mixed backgrounds, students should be engaged to think about potentially interesting project ideas. Each conducted project should include at least one participant with a (digital) humanities and one with a computer science background to expedite interdisciplinary exchange. This constraint generates different supervisory roles:

The Mediator: Projects involve students having a computer science and a humanities background alike. The entire project is managed by the students, whereas the teacher might supervise in the form of a mediator during meetings. While students face typical pitfalls of interdisciplinary projects, such projects still brought forth considerable results, one of which is shown in Figure 1.

The Real & the Fake Humanities Scholar: When the number of students with a humanities background is too low, project groups that only include computer science students need to be complemented with domain experts. For some of the projects, I was able to involve partners from the humanities, the real humanities scholars, with research interests targeted towards available data sets. This setting generated very good results (see Figure 2) and guaranteed the steepest learning curve for computer science students as they cooperated with domain experts experienced in digital humanities. However, for some projects I, educated in computer science, needed to act as a humanities scholar, being the least favorable setting as only fake interdisciplinary discussions are possible.

The Helper: On the other hand, the number of computer science students joining a course might be limited. In that case, I could advise students in the role of a computer scientist, better suitable considering my own background. The focus in such projects was rather on data modeling and acquisition as well as applying existing tools and libraries than developing new solutions. An example is shown in Figure 3.

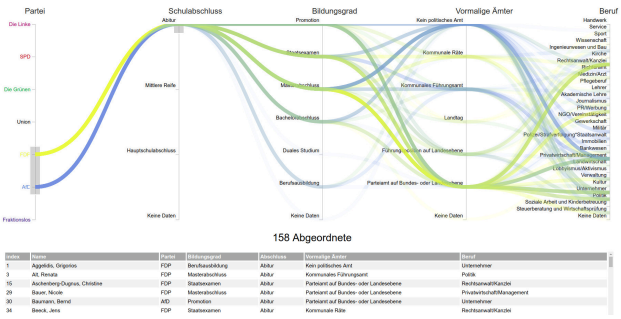


Figure 1: The mediator project developed a method to semi-automatically extract biographical information about members of the German Bundestag in 2019, and the adapted stream visualization allows multifaceted exploration of biographical features.

Figure 2: The real humanities scholar project focused on the development of an interactive tag cloud that supports composing engineering branches based on the study subjects of engineering professors. More information can be found in a related publication on the project (Meinecke and Jänicke, 2018).

Figure 3: The helper project focused on the contents published at three German websites known for publishing fake news articles. For comparatively analyzing the results, the TagPies visualization (Jänicke et al., 2018) was adapted.

I recommend making such courses accessible to all students as it prepares them best for potential future collaborations in a digital humanities context. More detailed information on theoretical contents, conducted student projects and course reflections can be found in my related IVAPP article (Jänicke, 2020).

References

Bastian, M., Heymann, S., Jacomy, M., et al. (2009). Gephi: an open source software for exploring and manipulating networks. ICWSM, 8:361-362.

Jänicke, S. (2016). Valuable Research for Visualization and Digital Humanities: A Balancing Act. Workshop on Visualization for the Digital Humanities, IEEE VIS 2016, Baltimore, Maryland, USA.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2017). Visual Text Analysis in Digital Humanities. In Computer Graphics Forum (Vol. 36, No. 6, pp. 226-250).

Jänicke, S., Blumenstein, J., Rücker, M., Zeckzer, D., & Scheuermann, G. (2018). TagPies: Comparative Visualization of Textual Data. In VISIGRAPP (3: IVAPP) (pp. 40-51).

Jänicke, S. (2020). Teaching on the Intersection of Visualization and Digital Humanities. In: 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (pp. 100-109). SCITEPRESS Digital Library.

Meinecke, C., & Jänicke, S. (2018). Visual Analysis of Engineers' Biographies and Engineering Branches. Leipzig Symposium on Visualization in Applications 2018 (LEVIA'18).

Munzner, T. (2014). Visualization Analysis and Design. A K Peters Visualization Series, CRC Press.

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Visual Languages, Proceedings, pages 336-343.

Sinclair, S., & Rockwell, G. (2020). Voyant-Tools. https://voyant-tools.org/

Windhager, F., Federico, P., Schreder, G., Glinka, K., Dörk, M., Miksch, S., & Mayr, E. (2018). Visualization of Cultural Heritage Collection Data: State of the Art and Future Challenges. IEEE Transactions on Visualization and Computer Graphics, 25(6), 2311-2330.

Fast Search with Poor OCR

Taivanbat Badamdorj
tbadamdorj@ualberta.ca
University of Alberta, Canada

Adiel Ben-Shalom
adiel.benshalom@gmail.com
Tel Aviv University, Israel

Nachum Dershowitz
nachumd@tauex.tau.ac.il
Tel Aviv University, Israel

Lior Wolf
liorwolf@gmail.com
Tel Aviv University, Israel Facebook AI Research

Keywords: Optical character recognition
Topics: Europe, English, 20th Century, information retrieval and querying algorithms and methods, optical character recognition and handwriting recognition, History, Central/Eastern European Studies

Introduction

Indexing and searching historical documents has garnered attention in recent years due to massive digitization efforts of important collections worldwide. Pure textual search in these corpora is problematic since optical character recognition (OCR) is infamous for performing poorly on such historical material. We propose a novel text-based method for searching through noisy text.

We implemented the proposed method for the Wiener Library, one of the most extensive archives on the Holocaust and Nazi era. Established in 1933, the library's unique collection includes press cuttings, eyewitness testimonies, photographs, as well as published and unpublished works from that era. It includes approximately 150,000 books, reference works, pamphlets and journals and over one million indexed newspaper clippings, unpublished memoirs, and interviews. These documents have been digitized and made available online. Hence, it's of great importance to have a fast search tool for them.

We present an easy-to-implement method that enables scholars and the public to effectively search such large collections of textual material that would otherwise be inaccessible due to the lack of reliable transcriptions. Its advantage is its simplicity, in contrast to existing image-based methods (such as various word-spotting methods [1, 2, 3, 4], which are technically involved and need to be rebuilt/retrained for each script style within a dataset. It is based on existing OCR tools that work well for a variety of typefaces and languages. The ease and low cost of implementation enables anyone to apply it to various datasets without substantial expertise in computer vision.

Method

Given a text query, we would like to find the correct matches among the noisy candidates output by the OCR. Our method consists of (A) preprocessing the images, (B) obtaining noisy candidate words using an OCR engine, (C) encoding the query and candidate words into vectors, (D) learning a common subspace between them, and finally (E) ranking the candidates according to distance from the query.

For preprocessing (step A), we remove black borders surrounding documents in the images and adjust contrast using the CLAHE algorithm [5] because many documents have non-uniform lighting.

Encouraged by recent improvements in OCR software, we chose the Tesseract engine [6, 7] (B) to obtain candidates. Tesseract is open-source and works well for a variety of typefaces and languages.

Next, we encode the candidates and queries in fixed-size vectors (C) using the recent pyramidal histogram of characters (PHOC) representation [1]. Words with similar spellings are close together in the PHOC vector space.

We then use canonical-correlation analysis (CCA) to learn a common subspace between noisy OCR readings and ground truth (D). This step performs "soft-correction" by projecting OCR misreadings closer to their ground truth texts in the learned subspace.

Lastly, we find the correct candidates using a nearest-neighbor search (E). The most commonly used distance metric for this is cosine distance. However, nearest-neighbor search is by its very nature asymmetric. This causes faulty matching in high-dimensional spaces [8]. We use the recently proposed cross-domain similarity local scaling (CSLS) [9] to mitigate this issue.

Experiments and Results

We tested the accuracy and speed of our system in an information-retrieval setting on a subset of the Wiener Library's "500 Document Collection," which makes up the main component of the original materials and comprises over 75,000 images that have been scanned from microfilm and microfiche copies of the originals. We annotated 18 pages, totalling 4284 words. Furthermore, we created 20 random splits, each split containing 9 pages for learning the common subspace using CCA, and 9 pages for testing.

We used a set of 96 Latin characters for the PHOC representations, thus resulting in a 1440 size binary vector for each word.

Method	PHOC (C)	CCA (D)	Distance (E)	mAP (%± s.d.)	Time (s)
Edit distance	✗	✗	Edit distance	**81.94** ± 4.56	90
CCA and CSLS	✓	✓	CSLS	81.85 ± 4.70	11
CSLS	✓	✗	CSLS	81.66 ± 4.73	10
CCA and Cosine	✓	✓	Cosine	81.36 ± 4.79	11
Cosine	✓	✗	Cosine	81.09 ± 4.76	**8**

Table 1: Search results, with and without PHOC encoding or CCA projection and using different metrics. Last column gives average time on a CPU to search through the labeled test set.

Pair	p-value	Reject H_0?
Edit distance ↔ CCA and CSLS	0.32	No
CCA and CSLS ↔ CSLS	1.44×10^{-4}	Yes
CSLS ↔ CCA and Cosine	2.06×10^{-6}	Yes
CCA and Cosine ↔ Cosine	2.15×10^{-6}	Yes

Table 2: Statistical significance test results.

We tested different versions of our system against edit distance. Table 1 shows the performance in the information-retrieval setting. The results of the paired t-tests can be seen in Table 2. Finally, the last column of Table 1 shows the timing comparisons between the methods. The best version of our system (CCA and CSLS) is 9 times faster than edit distance, and achieves almost the same mean average precision (mAP).

The difference in performance between edit distance and our method is statistically insignificant, although edit distance performs slightly better on average. The differences between all other methods are statistically significant. This means that the two additions of (1) learning a common subspace to perform soft correction (step D) and (2) using the CSLS metric (step E) were effective in improving the overall performance of our method.

Figure 1: Qualitative search results for the entire dataset. Bounding boxes in pink. Our model works wellwith faded ink and approximate matching.

Figure 1 shows qualitative search results through the entire dataset. The system finds very long words ("Nationalsozialistische"), as well as approximate matches to a query that has slightly different readings ("Rossenstrasse" vs. "Rossenstraße"). As can be seen, OCR performs well on a variety of fonts.

Conclusion

We presented a fast and accurate text-based search that is easy to implement and which requires minimal fine-tuning for any given setting. It appears to provide an excellent balance between speed and accuracy. It might make sense to use edit distance to rerank the top results.

We applied the system described here to all German-language documents in the Wiener collection. When embedded---as planned---in the library's search tool, this will provide WWII scholars a valuable tool to search effectively through these important historic documents.

State-of-the-art OCR works well out of the box for a variety of different languages and fonts. This is important for our dataset because it includes German, Polish, English, and Hebrew texts. Within each language, there are also a variety of styles that must be dealt with.The method does depend on reasonable, though imperfect, OCR results.We also experimented with word-spotting techniques, which could be combined with OCR when needed. See [10] for another example of leveraging poor OCR for finding related texts.

Our method is currently being applied to the other languages and to the much larger Yad Vashem collection.

Acknowledgements

We thank Alexey Pechorin and Zahi Hazan for technical help, Hila Buzaglo, Gila Michlovski, Naama Scheftelowitz, Roni Stauber, and their teams at the university libraries for providing the data and helping evaluate outcomes, and Leo Corry for making this all happen.

References

[1] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(12):2552-2566, 2014.

[2] Raghavan Manmatha, Chengfeng Han, Edward M. Riseman, and W. Bruce Croft. Indexing handwritingusing word matching. In First ACM International Conference on Digital Libraries, pages 151-159, 1996.

[3] Sebastian Sudholt and Gernot A. Fink. PHOCNet: A deep convolutional neural network for wordspotting in handwritten documents. In 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), pages 277-282. IEEE, 2016.

[4] Ismet Zeki Yalniz and Raghavan Manmatha. An efficient framework for searching text in noisy document images. In 10th IAPR International Workshop on Document Analysis Systems, pages 48-52. IEEE, 2012.

[5] Garima Yadav, Saurabh Maheshwari, and Anjali Agarwal. Contrast limited adaptive histogram equalization based enhancement for real time video system. In International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 2392-2397. IEEE, 2014.

[6] Ray Smith. An overview of the Tesseract OCR engine. In Ninth International Conference on Document Analysis and Recognition (ICDAR), volume 2, pages 629-633. IEEE, 2007.

[7] Ray Smith. Hybrid page layout analysis via tab-stop detection. In 10th International Conference on Document Analysis and Recognition (ICDAR), pages 241-245. IEEE, 2009.

[8] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research, 11:2487-2531, Sep 2010.

[9] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In 6th International

Conference on Learning Representations (ICLP, Vancouver, BC). OpenReview.net, April 2018.

[10] Nachum Dershowitz, Daniel Labenski, Adi Silberpfennig, Lior Wolf, and Yaron Tsur. Relating articles textually and visually. In 14th IAPR International Conference on Document Analysis and Recognition (ICDAR, Kyoto, Japan), pages 275-281, November 2017.

Mapping the Gay Guides: Understanding Queer Spaces in Pre- and Post-Stonewall America

Amanda Regan
aregan@smu.edu
Southern Methodist University, United States of America

Eric Gonzaba
egonzaba@fullerton.edu
California State University Fullerton, United States of America

Keywords: digital history, mapping, queer DH, gay bars
Topics: English, North America, 20th Century, spatial & spatio-temporal analysis, modeling and visualization, History

Mapping the Gay Guides is a project aimed at understanding often ignored queer geographies through an interactive web application and digital public history project. Drawing on and making accessible the Damron Guides, an early but longstanding travel guide aimed at gay men since the early 1960s, this online mapping project explores different dimensions of American gay life through time, from bars and nightlife, bookstores, cinemas, and churches. Utilizing digitized data and visualizations from the app, the website will present a historical analysis of the changing ways that gay spaces were defined. This presentation will discuss the initial-beta version of the project and our findings.

Accessibility and Reception Vector Semantics, Reading Publics, and the Changing Reception of Literary Works

Andrew Dunn
adunn@gradcenter.cuny.edu
The Graduate Center, CUNY, United States of America

Keywords: vector semantics, accessibility, reception theory
Topics: Europe, English, North America, 20th Century, Contemporary, cultural analytics, text mining and analysis, Cultural studies, Literary studies

This paper explores the possible applications of vector semantics in the branch of literary study known as Reception Theory. Specifically, this paper suggests that vector semantics can help us identify the shared "conceptual substrates" that identify and distinguish historically delineated publics; and, consequently, that vector-space models can enrich our understanding of a literary work's changing reception over time, in terms of its conceptual, semantic, and aesthetic accessibility.

Archival Justice in the DH Classroom: Examining El Diario de la Gente and the Chicanx Movement in Boulder, Colorado

Nickoal Eichmann-Kalwara
nickoal.eichmann@colorado.edu
University of Colorado Boulder, United States of America

Keywords: Pedagogy, Chicanx Movement
Topics: English, North America, 20th Century, curricular and pedagogical development and analysis, digital activism and advocacy, Chicano/a/x, Latino/a/x studies, History

This presentation addresses the role that archival justice can play in the digital humanities classroom,

as a means to not only enable exploration of methodological approaches in DH, from text mining and geospatial analysis, but also to address the disciplinary decorum of DH. In particular, archival justice centralizes postcolonial and feminist approaches to DH pedagogy and research, while playing a vital role in unearthing silenced histories. The author will share the pedagogical considerations and challenges of this approach to teaching a graduate-level DH course through analysing the independent Chicanx student newspaper, El Diario de la Gente, which was published 1972-1983 at the University of Colorado Boulder. The presentation will also highlight students' final projects that resulted from the class.

Mapping the Ghadr Directory: Dimensions of Diasporic Anticolonialism and State Surveillance

Eliza C. Bettinger
ecb4@cornell.edu
Cornell University, United States of America

S. Ani Mukherji
mukherji@hws.edu
Hobart and William Smith Colleges, United States of America

Keywords: surveillance, anticolonialism, data cleaning, maps, networks
Topics: Asia, Global, English, 20th Century, database creation, management, and analysis, spatial & spatio-temporal analysis, modeling and visualization, Geography and geo-humanities, History

This paper reports on a collaborative work-in-progress with two parts: (1) the creation of an open, reusable dataset from The Ghadr Directory, a 1917 British intelligence publication tracking the biographies of migrant South Asian anticolonial activists and (2) our effort to map the spatial and social dimensions of the global Ghadr network of using the dataset.

Each entry of the directory includes an activist's name, district of origin, locations of previous and current activity, and known associates. This information makes the directory a rich source for mapping a diasporic network of anticolonialist activism and the reach of British surveillance in the interwar period.

Entries in the directory vary in detail and length, presenting challenges in transformation to a standardized, analyzable format, and an opportunity to critically examine data "cleaning" and scale. We will use the resulting dataset to address questions of networks and movements among Ghadr activists.

Crossroads, shortcuts, detours, bypasses or dead ends? Attempts to standardization and interoperability within the context of the CLARIAH-DE project

Daniela Monika Schulz
schulz@hab.de
Herzog August Bibliothek Wolfenbüttel, Germany

Keywords: national research infrastructures, standardization, interoperability
Topics: Global, Europe, English, Contemporary, digital research infrastructures development and analysis, metadata standards, systems, and methods, Humanities computing

Within recent years, both the DARIAH-DE[1] as well as the CLARIN-D[2] project established themselves as digital research infrastructures for tools and research data in Germany, and also worked on developing materials for teaching and further education in the field of digital humanities[3]. Whereas DARIAH-DE had a clear focus on textual data and e.g. its representation in form of digital editions, the focus of CLARIN-D was more the provision of data and tools for linguistic analysis. Funded by the German Federal Ministry for Education and Research[4], both infrastructures will be merged within CLARIAH-DE (2019-2021), building on earlier cooperation and coordination processes, and aiming at interoperability in the future (also in accordance with the FAIR data principles[5]).

In CLARIAH-DE Work Package (WP) 1 "Research Data, Standards and Procedures"[6] one potential

solution to enable interoperability is the use of prevalent TEI customizations like the German Text Archive's base format (DTABf)[7] as an exchange or target format. While the DTABf is already established as a pivot format for text collections (especially for the annotation of full texts of historical prints, and for newspapers and simply structured manuscripts), it will be now evaluated to what extent the DTABf can be applied to the diverse field of digital scholarly editing. Although such an exchange or pivot format may not be able to represent all previously coded information in the original depth, the underlying idea is that it still can be understood as a core data set, and therefore function as a common denominator.

The evaluation is based on case studies of selected digital editions, and led by the following questions: Which phenomena in the digital editions are currently not covered by the DTABf tag set?Which phenomena are encoded differently in the DTABf than in the digital editions? Is a mapping more or less unproblematic or are there fundamental (semantic) differences?Which information is required by the DTABf, but has not been encoded in the digital editions?

The encoded phenomena can be classified as loss-free transferable, partially transferable and missing or lost. Depending on the effort and the outcome, a conversion might result in either a DTABf-valid text, in which all necessary information has been encoded accordingly, but in which the encoding is also limited to the elements and attribute-value pairs permitted in the DTABf, or in a DTABf-compliant text, which contains a core encoding according to the DTABf, but which may also have annotations going beyond the DTABf.

To ensure the diversity and thus the representability of the underlying data, the editions chosen to serve as case studies for the evaluation should differ significantly from each other. Therefore, seven criteria were developed based on a review of relevant literature and existing editions. These criteria, for example the editorial model applied, the TEI modules used or the source material, resulted in an edition matrix (EdMa), which - beyond the purposes of the CLARIAH-DE-project - should enable a rough categorisation of digital editions in general.

The paper aims at outlining and discussing the problems in harmonizing standards between the two former separated infrastructures and the actions that have been taken so far, to bypass these obstacles. By using CLARIAH-DE as a concrete example, the presentation will therefore contribute to the general discussion on standardization within the context of national research infrastructures.

[1] https://de.dariah.eu/en/startseite. All links have been accessed in June 2020.

[2] https://www.clarin-d.net/en/.

[3] List of services developed and/or maintained in the context of DARIAH-DE: https://de.dariah.eu/en/list-services, for CLARIN-D see https://www.clarin-d.net/en/.

[4] https://www.bmbf.de/.

[5] https://www.force11.org/group/fairgroup/fairprinciples.

[6] In the WP members of the following institutions are cooperating: Chair of Media Informatics (University of Bamberg), Berlin-Brandenburg Academy of Sciences and Humanities, Göttingen State and University Library, Leibniz Institute for the German Language (Mannheim), and the Herzog August Library Wolfenbüttel.

[7] http://www.deutschestextarchiv.de/doku/basisformat/.

DIALOGUES ON DETENTION - AN ITERATIVE, INCLUSIVE APPROACH TO 3D VISUALISATIONS OF COLONIAL CAMPS IN KENYA

Chao Maina
mainachao@gmail.com
African Digital Heritage, Kenya

Keywords: colonial, kenya, visualisation
Topics: Africa, English, 20th Century, digital research infrastructures development and analysis, digitization (2D & 3D), African and African American Studies, History

This paper explores the use of 3D visualisations and new media in increasing participation and engagement with colonial history in Kenya. Lying at the intersection of digital technologies, silenced histories and colonial legacy, the study explore ways of creating iterative, community centered approaches to 3D digital reconstructions.

Exploring Guangxu-era missionary activities in Taiwan from Chinese Recorder, Dan-Hsin Archives and Ming-Qing Taiwan Administrative Archives through DocuSky

I Mei Hung
Research Center for Digital Humanities, National Taiwan University

Chijui Hu
huchijui@gmail.com
Research Center for Digital Humanities, National Taiwan University

Jieh Hsiang
Research Center for Digital Humanities, National Taiwan University

Keywords: DocuSky, Chinese Recorder, Dan-Hsin Archives, Ming-Qing Taiwan Administrative Archives, digital humanities platform
Topics: Asia, English, 19th Century, 20th Century, database creation, management, and analysis, digital research infrastructures development and analysis, Asian studies, History

In this paper we use materials in two languages from three different archives and using the DH platform DocuSky to consolidate them. The three archives are Dan-Hsin Archives, Ming-Qing Taiwan Administrative Archives, both in Chinese, and the Chinese Recorder, in English. We use the missionary activities in Taiwan during Guangxu era(1875-1908) as an example.

We first retrieve documents of Guangxu era from these three sources and combine them into one corpus on DocuSky. The latter involves reorganizing the metadata, tags and connecting meaningful terms. Tools in DocuSky enable us to analyze the documents even from different sources and languages. We use two examples to show the efficacy of our method. First, we show the different perspectives from missionaries and Chinese during the Sino-French War(1885) revealed from the documents. We also show how DocuSky's DocuXML format can easily link to Palladio to show social network among foreign missionaries and native believers.

Semi-Automatic Identification of Travelogues

Jan Rörden
jan.roerden@ait.ac.at
AIT Austrian Institute of Technology, Austria

Rainer Simon
AIT Austrian Institute of Technology, Austria

Doris Gruber
Austrian Academy of Sciences, Austria

Martin Krickl
Austrian National Library, Austria

Bernhard Haslhofer
AIT Austrian Institute of Technology, Austria

Keywords: Travelogues, Digital Humanities, Machine Learning
Topics: Asia, Europe, English, 15th-17th Century, 18th Century, 19th Century, artificial intelligence and machine learning, bibliographic analysis, History, Informatics

Travel literature represents a rich source of information about the past, and has been of increasing interest in the scholarly community (c.f. Salzani & Tötösy de Zepetnek, 2010; Belgum et al., 2018). The *Travelogues*[1] project aims to study what we can learn from past views of foreign regions, cultures and religions in the light of present-day challenges such as mass tourism, migration and globalization. Comprising a team of historians, librarians and data scientists, *Travelogues* applies a transdisciplinary approach, combining quantitative and qualitative analytical methods to study a large-scale corpus of German language travelogues.

The project focuses on German-language holdings of the Austrian National Library printed between 1500 and 1876, including 167,570 digitized volumes. Those volumes have previously been digitized and processed by Optical Character Recognition (OCR) in the *Austrian Books Online* project-a public-private partnership of the Library and Google Books. In order to facilitate analysis of this vast and heterogenous collection, the project faces a number of challenges. The first challenge is to compile a corpus that includes as many travelogues from the inventory as possible. The second challenge is the profiling of the corpus at scale, analyzing it specifically for aspects of geographical coverage, salient terms over time and intertextuality. Finally, the key challenge lies in the identification of depictions of *otherness* in the corpus, and evolution of those depictions over time.

Travelogues is a work in progress. In this paper we will describe our results so far-how we created the corpus and approached the task of profiling-and present our plans for the upcoming steps required for a detailed analysis of the corpus.

As the intellectual basis for the project, historians first established a definition for this project's use of the term *travelogue*. In the context of the project, a travelogue is defined as a specific form of media that records the experiences of a factually undertaken journey. Applying this definition, we created a balanced ground truth of digitized travelogues and non-travelogues (works that could belong to any other genre). To account for variations in the data such as document length, OCR quality or orthographic differences, we created separate ground-truth datasets for different time periods: the 16th, 17th and 18th centuries and 1800-1876. This is a manual and time-consuming process involving several steps, including keyword and metadata searches of the collection, cleansing and enrichment of heterogeneous metadata and comparisons with both contemporary and modern travelogue bibliographies (e.g. Chatzipanagioti-Sangmeister, 2006; Griep & Luber, 1990; Treue, 2014; Yerasimos, 1991). Every book we identified using this method was independently verified by a historian and a librarian.

Based on the ground truth for each period, we trained and evaluated different machine learning algorithms to classify works as either travelogues or non-travelogues. This evaluation was done using five-fold cross-evaluation on a training set and a validation set. Using the best-performing approach, we applied the models for each time period and classified all documents not part of the ground truth (a total of 161,522 books). Our model returned a confidence score, essentially quantifying the likelihood that a given document is or includes a travelogue. The top 200 findings for each time period (800 in total) were then manually evaluated by our domain experts in order to confirm the validity of the automated results. Our process revealed 345 previously-unknown volumes of travelogues that were not listed as travelogues in any bibliography we consulted so far, nor could they be found using conventional metadata search methods (e.g. searching for different spelling variations of the German word for *travel)*. Although the 345 newly-discovered travelogues did not noticeably differ in their content from the previously-known canon, their materiality was particular. A large number of them were originally published as part of larger documents (such as serial publications, collected volumes or diaries) that, due to the lack of metadata in the library system, usually cannot be found with the traditional methods of the humanities as described above. Although we did not segment the documents into smaller entities (e.g. pages or chapters) for classification (Underwood et al., 2013), this shows that our methodology leads to robust results concerning documents that are only partially considered part of a genre, as in our case with travelogues. Additionally, we successfully proved that our methodology can expand traditional bibliographic research and help save time for domain experts. We have already described our classification task in detail (Rörden et al., 2020).

Our next steps concern the analysis and historical contextualization of the corpus. We are currently creating a searchable index of the entire document inventory (travelogues as well as non-travelogues) to enable exploratory searches. Key exploration scenarios include plotting the number of travelogues published over time, optionally while filtering by various facets (such as authors, publishers, keywords or catalogue subject classifications); and exploring salient terms that feature more prominently in travelogues over non-travelogues, or in travelogues of a particular period vs. in those of another. We have also begun to generate maps of travelogues' geographical coverage by performing Named Entity Recognition, and resolving coordinates against the GeoNames gazetteer.

Furthermore, we have taken the first steps to analyze intertextual relations between documents (e.g. Dörr & Kurwinkel, 2014; Rajewsky, 2002), experimenting with a mix of approaches including n-gram fingerprinting (c.f. Stein, 2007) and paragraph vectors for document representation (Le & Mikolov, 2014). This has been combined with clustering and text passage alignment using the BLAST-based text reuse algorithm developed by Vierthaler and Gelein (2019). By applying the algorithm corpus-wide, we hope to learn more about the relationships between the documents and their authors,

as well as how descriptions of and references to people, places and customs propagate through literature over extended time periods. Preliminary results seem promising, and have revealed what appear to be potential candidate cases of previously undocumented text-reuse. However, deeper analysis of the candidates and refinement of the methods are still ongoing. This method for the detection of intertextual relations is also a promising tool for clustering and relating works, not only based on literal title-strings but also on indexed full-texts, thus following the suggested implementation of the International Federation of Library Associations and Institutions Library Reference Model (IFLA LRM) into library catalogs (Decourselle et al., 2015; Rafferty 2015; Riva et al., 2017).

With our method, we were able to create the largest curated corpus of German-language travelogues to date-3,595 volumes, 345 of which were, to the best of our knowledge, not previously identified or findable as travelogues-thus proving that methods like ours can successfully expand the classic bibliographic methodology of the humanities and library sciences. We have made first steps towards enabling the interactive exploration of a number of relevant properties of the corpus. The next year will be dedicated to addressing the main research goals: the identification of intertextual relations between the travelogues in our corpus to deepen our understanding of how they depended on each other, and what this tells us about the circulation of knowledge, stereotypes and prejudices. This will ultimately lead to the question of how notions of *otherness* were depicted, how and why they changed over time, and what conclusions this allows concerning today's perceptions and biases. We have already published large parts of our corpus under a Creative Commons license.[2] Beyond the goals of our own project, we feel that the open availability of the corpus marks a significant contribution to the research community at large, and will invite further scholarship and collaboration around this exciting resource.

Acknowledgments

The work in the Travelogues project (http://www.travelogues-project.info) is funded through an international project grant by the Austrian Science Fund (FWF, Austria: I 3795) and the German Research Foundation (DFG, Germany: 398697847).

References

Belgum, Kirsten et al. "Mapping Travel Writing: A Digital Humanities Project to Visualise Change in Nineteenth-Century Published Travel Texts." Studies in Travel Writing, vol. 22, no. 3, July 2018, 306-24.

Chatzipanagioti-Sangmeister, Julia. Griechenland, Zypern, Balkan und Levante. Eine kommentierte Bibliographie der Reiseliteratur des 18. Jahrhunderts. 2 vols., Eutin: Lumpeter & Lasel 2006.

Decourselle Joffrey et al. "A Survey of FRBRization Techniques." Kapidakis Sarantos et al. (eds.). Research and Advanced Technology for Digital Libraries. TPDL 2015. Cham: Springer International Publishing 2015, 185-196.

Dörr, Volker C., Tobias Kurwinkel, editors. Intertextualität, Intermedialität, Transmedialität. Zur Beziehung zwischen Literatur und anderen Medien. Würzburg: Königshausen & Neumann 2014.

Le, Quoc, Tomas Mikolov. "Distributed representations of sentences and documents." Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML'14), vol. 32, no. 2, 2014, 1188-96.

Rafferty, Pauline. "FRBR, Information, and Intertextuality."Library Trends, vol. 63, no. 3, 2015, 487-511.

Rajewsky, Irina O. Intermedialität. Tübingen, Basel: Francke 2002.

Riva, Pat et al. IFLA Library Reference Model. A Conceptual Model for Bibliographic Information, edited by the International Federation of Library Associations and Institutions. Den Haag: IFLA 2017.

Rörden, Jan et al. "Identifying Historical Travelogues in Large Text Corpora Using Machine Learning." Anneli Sundqvist et al. (eds.). Sustainable Digital Communities. 15th International Conference, iConference 2020. Boras, Sweden, March 23-26, 2020. Proceedings, Cham: Springer International Publishing 2020, 801-15.

Salzani, Carlo, Steven Tötösy de Zepetnek. "Bibliography for Work in Travel Studies." Library Series, CLCWeb: Comparative Literature and Culture, July 2010 (updated on 14.09.2016), https://docs.lib.purdue.edu/clcweblibrary/travelstudiesbibliography.

Stein, Benno. "Principles of hash-based text retrieval." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval 2007 (SIGIR '07). New York: ACM 2007, 527-34.

Treue, Wolfgang. Abenteuer und Anerkennung. Reisende und Gereiste in Spätmittelalter und Frühneuzeit (1400-1700). Paderborn: Ferdinand Schöningh 2014.

Underwood, Ted et al. "Mapping Mutable Genres in Structurally Complex Volumes." Proceedings of the 2013 IEEE International Conference on Big Data, December 2013, 95-103.

Vierthaler, Paul, Mees Gelein. "A BLAST-based, Language-agnostic Text Reuse Algorithm with a MARKUS Implementation and Sequence Alignment Optimized for Large Chinese Corpora." Journal of

Cultural Analytics, March 2019, DOI: 10.22148/16.034, Dataverse DOI: 10.7910/DVN/2YYJ2B.

Yerasimos, Stephane. Les voyageurs dans l'Empire Ottoman (XIVe-XVIe siècles). Bibliographie, itinéraires et inventaire des lieux habités. Ankara: Imprimerie de la Société Turque d'Histoire 1991.

[1] https://travelogues-project.info/.

[2] http://github.com/travelogues/travelogues-corpus.

Modeling possible stories. A digital edition of Giovanni Domenico Tiepolo's Divertimento per li Regazzi

Rostislav Tumanov
rostislav.tumanov@ikg.uni-stuttgart.de
University of Stuttgart

Gabriel Viehhauser
viehhauser@ilw.uni-stuttgart.de
University of Stuttgart

Alina Feldmann
University of Stuttgart

Barbara Koller
University of Stuttgart

Keywords: digital art history, CIDOC, digital edition, Tiepolo
Topics: Europe, English, 18th Century, data modeling, scholarly editing and editions development, analysis, and methods, Art history, Literary studies

Modeling possible stories. A digital edition of Giovanni DomenicoTiepolo's Divertimento per li Regazzi

Abstract

The paper presents a digital edition of Giovanii Domenico Tiepolo's drawing series *Divertimento per li Regazzi*. The series depict *Pulcinella*, a well-known character from the *Commedia dell'Arte* in different situations and can be interpreted as kind of a life story of the character. However, unlike in 'normal' narratives, the single drawings can not be ordered in one linear sequence but offer the recipient a variety of possibilities to arrange different stories. To model and visualize this open structure we draw on the CIDOC conceptual reference model and discuss a possible display for the front-end of the edition. Thus, the main contribution of the paper lies in the exploration of new ways for a digital presentation of narrative drawings series at the intersection of art history and narratology

Introduction

The drawing series *Divertimento per li Regazzi* created by the Venetian artist Giovanni Domenico Tiepolo in the late 18th century lies at the intersection of narrative and figurative art: It features scenes that depict Pulcinella, the well-known character from the *Commedia dell'Arte*, in what appears to be different stages of his life, his birth, his adolescence or his death. Thus, the series can be 'read' as a kind of life story of Pulcinella. However, other than in linear narratives, the exact sequence of the drawings is unclear and even inconsistent. Furthermore, it is subverted by a playful arrangement of motifs that establish paradigmatic relationships between single drawings of the series or interlink them to works of other artists.This richness of possible relations is hard to grasp with conventional means: In the printed catalogue by Adelheid Gealt[1], which still builds the base for scholarly research on the series, the drawings are - necessarily - presented in a static sequence that suggests a specific order of the pictures. As current endeavors in the field of textual scholarship have shown, such limitations of the printed media can be overcome with the help of digital editions that are able to present material in a dynamic way[2] [3]. Thus, a transfer of these methods on figurative objects seems promising.In our paper we want to report on our digital edition of the Divertimento, which takes into account the potential openness of the series and at the same time shows itsinterconnectedness. For this purpose, we establish a data model that draws on the CIDOC Conceptual Reference Model [4], which offers a solid, yet variable enough framework. Furthermore, we suggest a workflow that includes the conceptualization of a frontend that meets the requirements of a flexible display of the Divertimento. Therefore, the main contribution of our paper lies in the combination of different methods to establish new ways to present material at the intersection of textual and figurative art.

Divertimento per li Regazzi

The Divertimento consists of 104 drawings that depict Pulcinella in different situations. The individual Pulcinella-figures show enough resemblance to suggest that the series can be understood as a story of Pulcinella's life from the cradle to the grave. However, because of numerous inconsistencies it is

not possible to interconnect all the drawings into one concise narrative[5]. E.g., there are two drawings that show Pulcinella getting executed in two different ways and even one more drawing that shows his pardon. Thus, in contrast to 'normal' narratives, there is no clearly intended arrangement of a single plot. Instead, the series offers a wide range of possible stories, to be constructed by the recipient. Finding a way to handle these possibilities demands the application of specific perspectives that arrange the material in a manageable manner. E.g., from a narratological point of view, not all of the possible arrangements appear to be reasonable: Biographic readings of the series would suggest that drawings, which show Pulcinella as a child, should be sorted before pictures that show him as grown-up.

Furthermore, Pulcinella's life story is supplemented by a paradigmatic dimension that is established by the repetition of specific motifs (e.g, several drawings contain a pot of gnocchi that suggests relations between those drawings). And finally, these internalparadigmatic relations are supplemented by external links to other works of art, motifs or practices. The scene that depicts the shooting of Pulcinella, e.g., clearly refers to the *Grandes Misères de la guerre of Jacques Callot*, a contextualization, which sheds new light on the drawings.

This structure leads to several requirements for our edition:- The goal of the edition should not be to reconstruct a single linear sequence, but rather to show possible links that are suggested by the series and enable the recipient to find diverse ways through the narrative space of the Divertimento.- To achieve this goal, a data model is needed that provides a logical narrative framework without determining a fixed sequence.- The data model has to take into account syntagmatic as well as paradigmatic relations.- It should be capable to grasp internal as well as external references that reach out into the semantic web.

Workflow

To express the possible relations between the drawings we employ a graph data model by using RDF-triples and the CIDOC Conceptual Reference Model. More concretely, we draw on the OWL-based version of the CIDOC, the Erlangen-CRM[6], which we adapted with the help of the ontology-editor Protégé[7]. The triples are stored in a graph database (GraphDB[8]).We use the ARC2[9] framework to process SPARQL-queries in php and transform the results in HTML. To establish a link to Linked Open Data resources we use authority files like the iconclass classification system[10].

Modeling the data in CIDOC-CRM

A core concept of our approach is the modeling of our data in CIDOC-CRM, a top level ontology widely used in different contexts of the Digital Humanities [11]. We stick to the CIDOC-model not only for standardization purposes, but also because its event-centered structure is particularly apt for our goals. With the help of CIDOC, a flexible chronological structure can be established, which does not have to stick to a fixed sequence of single artefacts, but rather uses the notion of a 'virtual' life story to group the drawings of the **Divertimento**. Thus, the whole life story of Pulcinella (E4_Period) is segmented into subgroups of E5_Events (e.g., Pulcinella's birth, childhood, or death) and E7_Activities (which further specify the events). E5_Events and E7_Activities are brought into a chronological order by using the relation P120_occurs_before.Figure 1 shows the two red E5_Event-nodes 'Pulcinella's birth and childhood' and'Pulcinella's professions' and their connections. Yellow nodes represent the single drawings that are assigned to the E5_Events (note that some drawings like image 8, "the young Pulcinella observes the laborers", can be assorted to two events). The two E5_Event-nodes are connected by various relations (amongst them P120_occurs_before) through E7_Activity-nodes (colored in light blue) that subdivide the events further. Both E5_Eventnodes are also connected to the overall E4_Period-node 'Pulcinella's life story' in purple.

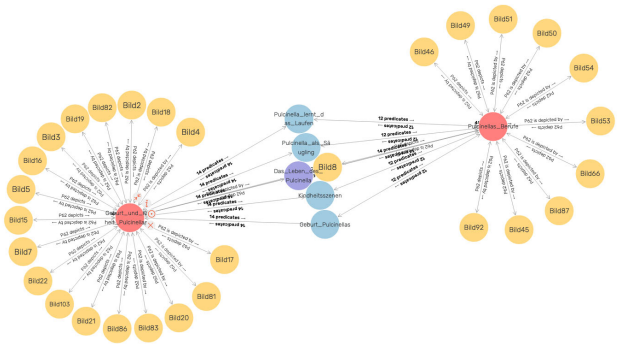


Figure 1: Detail from a network visualization of the data model, established with [8] (source: own)

Front end

Figure 2 shows a prototype of the front end for the edition. It focuses on a single drawing in the middle and presents possible following events in a carousel in the right column, where the respective drawings can be selected to add up to an individual 'story'. An overview of this story is presented in thumbnails in the left column.The featured motifs of the drawing are listed beneath and can be selected to jump tointernal or external pictures that hold the same motifs,

thus employing the paradigmatic relationships. Since all motifs are stored with their respective iconclass-ID, a further contextualization with resources of the semantic web can be achieved.

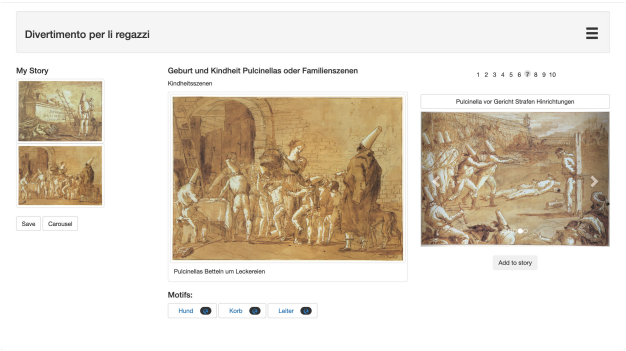


Figure 2: Prototype of the front end

References:

[1] Gealt, Adelheid M. (Ed.) (1986): Domenico Tiepolo. The Punchinello Drawings. New York: George Braziller.

[2] Sahle, Patrick (2013): Digitale Editionsformen, Zum Umgang mit der Überlieferun unter den Bedingungen des Medienwandels. Norderstedt: Books on Demand.

[3] Pierazzo, Elena (2016): Modelling Digital Scholarly Editing: From Plato to Heraclitus, in: Driscoll, Matthew James and Pierazzo, Elena (eds.): Digital Scholarly Editing. Theories and Practices. Cambridge, UK: Open Book Publishers, 41-58.

[4] Le Boeuf, Patrick / Doerr, Martin / Ore, Christian Emil / Stead, Stephen et al. (2019): "Definition of the CIDOCConceptual Reference Model", Version 6.2.6 May 2019, http://www.cidoc-crm.org/sites/default/file/CIDOC%20CRM_v6.2.6_Definition_esIP.pdf

[5] Vetrocq, Marcia (1979): "Domenico Tiepolo and the Figure of Punchinello", in: Gealt, Adelheid M. (eds..): Domenico Tiepolo's Punchinello drawings. Bloomington: Indiana University Art Museum, 19-20.

[6] http://erlangen- crm.org/

[7] https://protege.stanford.edu/

[8] http:// graphdb.ontotext.com/

[9] https://github.com/semsol/ arc2

[10] http://www.iconclass.org/

[11] Eide, Øyvind/ Ore, Christian-Emil Smith (2018): "Ontologies and data modeling", in: Flanders, Julia/Jannidis, Fotis (eds): The Shape of Data in Digital Humanities. Modeling Texts and Text-based Resources. London: Routledge 178-196.

Doing Ecological Identity Ecoparadox in Governor Babatunde Fashola's Speeches

Ayodeji Ayoola Adedara
ayodejiadedara@gmail.com
University of Lagos, Nigeria

Keywords: Identity, Eco-critical, Speeches, Ecoparadox, WordSmith.
Topics: Africa, English, Contemporary, eco-criticism and environmental analysis, rhetorical analysis, Linguistics

This study combines WordSmith Tools 7.0 with qualitative content analysis in investigating purposively selected gubernatorial speeches within the framework of eco-critical discourse analysis. It covers the eight-year tenure of a political actor in an emergent sub-Saharan Africa democracy that was deemed to have been one of the more 'progressive' since the advent of Nigeria's Fourth Republic in 1999. Covering the period 29 May 2007 to 29 May 2015, the study marks a longitudinal ecological analysis of discourse that aims to identify to what extent the government was ecocentric by highlighting the discursive strategies adopted in the speeches for creating a positive ecological identity for both the governor and his executive cabinet. The study represents use of digital methodology in sub-Saharan Africa, particularly in Nigeria, where a Centre for Digital Humanities University of Lagos (CDHUL) has just emerged under the auspices of the Alexander von Humboldt Foundation.

Data Art, Public Engagement, and Luso-Brazilian Literature and Culture Reflections on the Bird-watching Exhibition

Joshua Alma Enslen
joshua.enslen@westpoint.edu
West Point, United States of America

Keywords: Luso-Brazilian Literature and Culture, Public Humanities, Data Art, Text Mining, Network Analysis
Topics: South America, Comparative (2 or more geographical areas), Europe, English, 19th Century, 20th Century, Contemporary, mixed-media analysis, text mining and analysis, Cultural studies, Literary studies

In line with DH2020's focus on crossing paths, especially disciplinary ones, representations of indigenous (and other marginalized) perspectives in Brazil, and the possibilities of engagement proposed by public digital humanities initiatives, this paper will discuss the challenges and opportunities presented by the Bird-watching exhibition, an interdisciplinary and artisanal middle-ground where museum-goers were invited to critically examine Brazilian society through the lens of its most imitated poem.

"On the record": transcribing and valorizing qualitative interviews with XML-TEI

Florian Cafiero
florian.cafiero@gmail.com
CNRS / Sorbonne Université, GEMASS, France

Marie Puren
marie.puren@cnrs.fr
CNRS, Laboratoire de recherche historique Rhône Alpes, France

Keywords: TEI, qualitative methods, interview, spoken materials
Topics: Global, English, 20th Century, Contemporary, annotation structures, systems, and methods, text encoding and markup language creation, deployment, and analysis, Humanities computing

Qualitative interviews constitute an important research tool for disciplines such as history, sociology, ethnology or political science. Yet, despite rare initiatives (Cadorel et al., 2018), transcriptions are scarcely shared with other researchers. And their annotation is most of the time done only for a personal use, without following any sort of standard, and not meant to be shown to anyone.

In this paper, we advocate for the necessity of a more open management of these resources, and present a proposition for a XML-TEI-conformant standard, allowing for their accurate transcription and annotation. The ODD we present is aimed at facilitating systematic analyses of corpora of interview transcriptions, as well as at ensuring a better dissemination and re-usability of these resources. We rely as much as possible on existing TEI elements, but introduce a new element and a new attribute, to address the specificities of this kind of materials.

Interviews: a precious resource

Figuratively and literally, interview transcriptions are a precious resource. Producing them comes at a high cost: researchers must dedicate a lot of time and money to organize the interviews, travel, speak with the interviewees and transcribe the interviews. They should thus be used and re-used to their fullest potential. Finding a way to properly encoding them will help further qualitative or quantitative analyses by the researcher as well as by other colleagues taking interest in the source. Sharing effectively these resources would allow for comparisons between results obtained by researchers from various disciplines, at different periods and places, or to build larger corpora to obtain new results.

Addressing the reproducibility crisis

Human and social sciences have been targeted by many critics during the "replication crisis" (Ioannidis, 2005 ; Camerer et al., 2018) controversy. Research relying on qualitative analyses are not easily subject to the same reproducibility assessment, but should in many cases allow for "comparative re-production" (Markee, 2017): in a similar context, and following the same principles and questioning, will a new interview lead to results comparable to the ones previously made? Making one's transcriptions and annotations available would be a way to make this possible. It would also ensure that conclusions drawn from an interview are trusted, as a reviewer, colleague, or reader in general, could access the annotations underlying the researcher's analyses.

A standard adaptable by each research community

We propose to create an ODD (One Document Does It All), setting out which TEI elements and associated attributes can be used and in which context, and documenting our choices to future users. An ODD also enables to add new elements and attributes. And within a given community, it is possible to agree on an available ODD customization that will ensure the interoperability, shareability and reusability of the TEI files. To create this ODD, we mostly combine the elements and attributes

declared by the modules "Transcription of Speech" and "Language Corpora", and propose to add one new element and one new attribute.

Ethics and respect of privacy

A key concern in sharing qualitative interviews should be the respect of legal constraints and ethics principles. We thus propose the creation of a new element, to annotate passages that could not be freely shared. This element allows for a description of the deleted passage and the reason for its deletion. It is meant to ensure the protection of interviewees, while concealing as little relevant information as possible.

The creation of this new element relies on TEI best practices, and draws on the use of the element , employed to encode the damages done to a primary sources - for example by indicating that some text is lacking, and by supplying the lacking part with alternate text. The new element works in the same way to mention that a passage has been deleted and why it has been deleted. As well as the element, bears attributes such as "unit", "quantity" and "extent" to precise the length of the deletion. The reason of the deletion could then be expressed within an attribute "reason", more suitable in this context than the attribute "agent" born by . An element enclosed in the element gives also more information on the causes of the deletion (legal reasons, ethics code, personal moral judgement). We also propose an alternate encoding strategy by enclosing a element within to provide more information on the deletion. Moreover if the transcriber wishes to replace the deleted passage - e.g., replacing the name of a person with a pseudonym for ease of reading -, he or she may also use a element enclosed within .

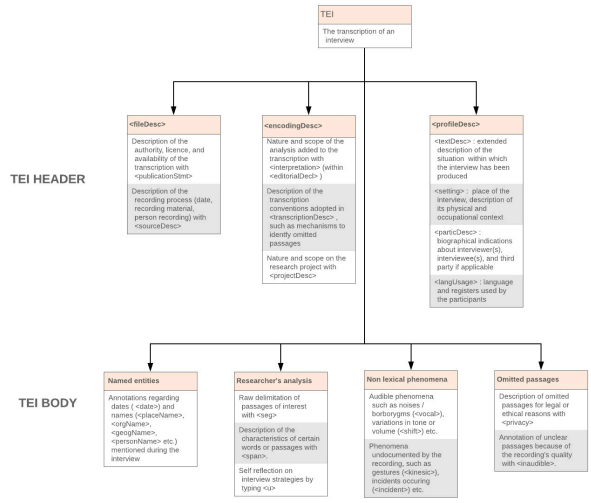
Reflecting on one's interview practices

The element is used to encode the different parts of speech given by the interviewers and the interviewees, with a "who" attribute to express who is the speaker. But qualitative interviews are not ordinary conversations: they are prepared by a researcher, implementing a strategy to get as much information as possible on a topic of interest. It is thus crucial to encode the researcher's comments on its own speech (Beaud, 1996): was the question prepared? spontaneous? what was its purpose (changing the subject/knowing more/ confirming a previous statement etc.) ? This is why we propose to add a "type" attribute to the list of already existing attributes born by , describing this kind of information.

Annotating the interview: sharing one's interpretation

In addition to simple content annotations (persons or places cited, dates evoked etc.), our model offers the possibility of sharing one's interpretations about relevant passages of the transcription.

We propose to use the element bearing an "xml:id" attribute to delimit the parts of the speech that are in need of further analyses. These analyses can then be provided via a element bearing a "target" attribute to identify which element is concerned, and a "type" attribute to express the nature of this analyses. The element may be used in conjunction with the element, but is more suitable to identify various parts of speeches under unique conceptual categories. Associating identified parts of speeches and specific conceptual categories is easy with TEI pointer mechanisms: for example, an "ana" attribute born by a element enables to associate this element with an element bearing an "xml:id" attribute.



References

BEAUD, Stéphane. L'usage de l'entretien en sciences sociales. Plaidoyer pour l'« entretien ethnographique ». Politix. Revue des sciences sociales du politique, 1996, vol. 9, no 35, p. 226-257.

BODARD, Gabriel. EpiDoc: Epigraphic documents in XML for publication and interchange. Latin On Stone: epigraphic research and electronic archives, 2010, p. 101-18.

CADOREL, Sarah, GARCIA, Guillaume, FROMONT, Emilie, et al. beQuali: Une Plateforme d'Archives Numériques en Sciences Sociales. In : Proceedings of the 1st International Conference on Digital Tools & Uses Congress. ACM, 2018. p. 23.

CAMERER, Colin F., DREBER, Anna, HOLZMEISTER, Felix, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nature Human Behaviour, 2018, vol. 2, no 9, p. 637.

IOANNIDIS, John PA. Why most published research findings are false. PLoS medicine, 2005, vol. 2, no 8, p. e124.

MARKEE, Numa. Are replication studies possible in qualitative second/foreign language classroom research? A call for comparative re-production research. Language Teaching, 2017, vol. 50, no 3, p. 367-383.

MIGUEL, Edward, CAMERER, Colin, CASEY, Katherine, et al. Promoting transparency in social science research. Science, 2014, vol. 343, no 6166, p. 30-31.

The TEI Guidelines, [online]. [Accessed 22 October 2019]. Available from: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Crowdsourcing Training Data Efficacy and Ethics

Alison Hedley
afhedley@gmail.com
McGill University, Canada

Keywords: crowdsourcing, machine learning, image analytics, data ethics, labour
Topics: Global, English, 19th Century, Contemporary, artificial intelligence and machine learning, crowdsourcing, Computer science, Humanities computing

Paid crowdsourcing presents a useful tool for building large-scale datasets of many kinds, including for humanities and cultural heritage work, but use of crowd labour is not without logistical and ethical challenges. This paper summarizes how the Visibility of Knowledge Project is using Mechanical Turk to develop training data and relates our best practices to broader challenges in paid crowdsourcing's ethics and efficacy. The experience of the VOK team suggests that devising itinerant communication tactics is necessary for any digital research projects that wish to use paid training data crowdsourcing in a manner that is both effective and ethical. Unfortunately, the nature of crowdsourcing work and paid platform design are such that the ethics of using crowd-labeled training data will almost certainly remain fraught, even as the need for large training datasets increases in many knowledge fields.

Foundations of Distant Reading. Historical Roots, Conceptual Development and Theoretical Assumptions around Computational Approaches to Literary Texts

Christof Schöch
schoech@uni-trier.de
University of Trier, Germany

Maciej Eder
Institute for Polish Language, Poland

Rosario Arias
University of Málaga, Spain

Pieter Francois
University of Oxford, UK

Antonija Primorac
University of Rijeka, Croatia

Keywords: Distant Reading, History of DH, Computational Literary Studies
Topics: Global, English, Contemporary, text mining and analysis, Literary studies

Introduction

The term 'distant reading' resonates across DH: It is played on in book titles (Distant Horizons, Underwood 2019) and adapted to new fields ('Distant Viewing', Arnold and Tilton 2019). It spurs alternative formulations ('Scalable Reading', Mueller 2012) and is present in mainstream media ("What is Distant Reading?", Schulz 2011). It is a popular and integrating term, but can take very specific meaning as well.[1]

However, the semantic content carried over in each case of adoption or adaption is often unclear. Recent debates, like the special issue of PMLA (On Franco

Moretti's Distant Reading 2017) or the paper by Nan Z. Da (Da 2019) and the reactions to it, have challenged some of the assumptions of 'distant reading'. Also, the polysemy of the term may have contributed to misunderstandings in these debates.

Therefore, our aim is to recover the historicity of the term 'distant reading', first introduced by Franco Moretti (2000) in his discussion of world literature as a system, by delineating how its meaning has changed over time and reconstructing some of the key theoretical assumptions it carries both as a term, a concept and a practice.

Historical roots

The pre-history to the concept now covered by the term 'distant reading' reaches back to the 15th century, when a rhetorical topos of "too many books" appeared (see Blair 2011). The solution was in excerpts and encyclopedias, based on the principles of compilation and summarization. The goal was to provide access to the essence of all relevant books instead of having to see them all at the same time. Of course, quantitative approaches to literary texts have appeared before the advent of computing (e.g. Mendenhall 1887) and computational approaches have diversified before the term 'distant reading' appeared (e.g. Ellegård 1962, Mosteller and Wallace 1963, Burrows 1987; see Hockey 2000).

Conceptual Development

When Franco Moretti first coined the term 'distant reading' in 2000, he used it with a meaning reminiscent of the compilatory origins of the concept, similar to "second-hand reading": using research literature, metadata or other short-cuts like titles and subtitles instead of reading the full text. From this starting point, and in parallel with more computational and more quantitative practices, Distant Reading has evolved to designate any computational, but especially quantitative, method of literary text analysis - so much so that the term now 'self-evidently implies computation' (Goldstone 2017, 637; see also Underwood 2017 and Bode 2017).

Theoretical Assumptions

A fundamental assumption of the earlier concept of 'distant reading' was that because metadata or secondary literature are created by humans who have read the full texts, they can stand in for the full text. Also, that the bird's eye's view provides insight into the longue durée and into literature as a system (Oberhelman 2015). A fundamental assumption of current Distant Reading research is that useful (even if imperfect) formal and quantifiable textual features can be used as indicators or proxies for relevant literary phenomena, hence the centrality of modeling (see McCarty 2005; Flanders and Jannidis 2019) in Distant Reading research practice. Finally, the idea that despite the broadening meaning of the term "literature" (decanonization), literary texts have a specific way of functioning that requires the adaptation of methods to this domain.

Conclusion

We hope that by more usefully contextualizing the development of the strategic term 'distant reading', we can help avoid misunderstandings in current debates about computational approaches in humanistic inquiry.

Bibliography

Arnold, Taylor, and Lauren Tilton. 2019. "Distant Viewing: Analysing Large Visual Corpora." Digital Scholarship in the Humanities. https://doi.org/10.1093/digitalsh/fqz013.

Blair, Ann M. 2011. Too Much to Know: Managing Scholarly Information before the Modern Age. New Haven: Yale University Press.

Bode, Katherine. 2017. "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History." Modern Languages Quarterly 78 (1): 77-106. https://doi.org/10.1215/00267929-3699787.

Burrows, John. 1987. Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Clarendon Press.

Da, Nan Z. 2019. "The Computational Case Agains Computational Literary Studies." Critical Inquiry.

Ellegård, Alvar. 1962. A Statistical Method for Determining Authorship: The Junius Letters, 1769-1772. Gothenburg: University of Gothenburg.

Flanders, Julia, and Fotis Jannidis, eds. 2019. "The Shape of Data in the Digital Humanities: Modeling Texts and Text-Based Resources." Digital Research in the Arts and Humanities. London & New York: Routledge.

Goldstone, Andrew. 2017. "The Doxa of Reading." PMLA.

Hockey, Susan. 2000. Electronic Texts in the Humanities: Principles and Practice. Oxford: Oxford University Press.

McCarty, Willard. 2005. Humanities Computing. New York: Palgrave Macmillan.

Mendenhall, Thomas C. 1887. "The Characteristic Curves of Composition." Science ns9 (2145): 237-46.

Moretti, Franco. 2000. "Conjectures on World Literature." New Left Review, no. 1.

Mosteller, Frederick, and David L. Wallace. 1963. "Inference in an Authorship Problem." Journal of the American Statistical Association 58 (302): 275-309.

Mueller, Martin. 2012. "Scalable Reading." 2012. https://scalablereading.northwestern.edu/?page_id=22.

Oberhelman, David D. 2015. "Distant Reading, Computational Stylistics, and Corpus Linguistics:

The Critical Theory of Digital Humanities for Literature Subject Librarians." Digital Humanities in the Library: Challenges and Opportunities for Subject Specialists. Chicago: Illinois: Association of College, Research Libraries. https://shareok.org/handle/11244/33193?show=full.

"On Franco Moretti's Distant Reading." 2017. Publications of the Modern Language Association (PMLA) 132 (3): 613-89.

Schulz, Kathryn. 2011. "The Mechanic Muse - What Is Distant Reading?" The New York Times. https://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html.

Underwood, Ted. 2017. "A Genealogy of Distant Reading." Digital Humanities Quarterly 11 (2). http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Underwood, Ted. 2019. Distant Horizons: Digital Evidence and Literary Change. Chicago: The University of Chicago Press.

Notes

[1] This contribution has emerged from work in the COST Action "Distant Reading for European Literary History" (CA16204, https://distant-reading.net), a pan-European, collaborative networking project launched in 2017.

New Scholarship in the Digital Age Making, Publishing, Maintaining, and Preserving Non-Traditional Scholarly Objects

Matt Burton
University of Pittsburgh

Matthew J. Lavin
lavin@pitt.edu
University of Pittsburgh

Jessica Otis
George Mason University

Scott B. Weingart
Carnegie Mellon University, United States of America

Keywords: infrastructure, containerization, survey, digital publication
Topics: Europe, English, North America, Contemporary, digital archiving, digital publishing projects, systems, and methods, Book and print history

We discuss the myriad ways digital scholarship is being conceived, produced, distributed, and preserved in the digital humanities. This short paper is based on the results of an 18-month A.W. Mellon-funded project called Digits: A Platform to Facilitate the Production of Digital Scholarship, in which we interviewed 75 subjects involved in the creation, publication, maintenance, and preservation of Non-Traditional Scholarly Objects (NTSOs). We further studied containerization and other possible sociotechnical interventions suggested in secondary literature, which might reduce points of friction around NTSOs.

Counting Burned Books: A Quantitative Analysis of the Great Fire of London

Mark J. Hill
mark.john.hill@gmail.com
University of Helsinki, Finland

Keywords: Bibliographic analysis, social network analysis, metadata, seventeenth century history
Topics: Europe, English, 15th-17th Century, 18th Century, bibliographic analysis, network analysis and graphs theory and application, Book and print history, History

A quantitative overview of the Great Fire of London (1666) and its impact on the English book trade, both in London and more broadly. Data used includes the ESTC. Methods used include various statistical analyses; social network analysis; and traditional historical research.

Reading Like Alice Tricks of Perspective in Reading and Visualization

Stefania Forlini
sforlini@ucalgary.ca
University of Calgary, Canada

Uta Hinrichs
University of St. Andrews, Scotland

Keywords: reading; visualization; critical theory; literary studies;
Topics: EuropeEnglishNorth America19th Century, Contemporary, Interface design, development, and analysis, meta-criticism (reflections on digital humanities and humanities computing), Design studies, Literary studies

In the context of the complex intersection of technological changes, textual remediations, and changing modes of reading (highlighted by PMLA's recent two-part special issue on "Cultures of Reading"), this article argues for a more generous approach to reading that refuses to "sidestep" critical theory and its traditions of critical hermeneutics. Drawing from our work with humanistic visualization design and literary interpretation of science-fictional practices in Victorian periodicals, we propose an "Alice-in-Wonderland" approach to reading literary texts and collections. The emphasis is on harnessing the affordances and limitations of technological changes, textual remediations, and different modes of reading to reconnect us to the rich strangeness of our current moment.

De-centering Darwin opening up nineteenth century scientific correspondence and putting Darwin in context

Elizabeth Lauren Smith
els47@cam.ac.uk
Darwin Correspondence Project, United Kingdom

Keywords: Darwin, correspondence, TEI, science, letters
Topics: Europe, English, North America, 19th Century, public humanities collaborations and methods, text encoding and markup language creation, deployment, and analysis, History, History of science

This paper presents the Darwin Correspondence Project's work with Cambridge University Digital Library on Epsilon (Epistles of Science in the Long Nineteenth century). Following the work of Early Modern Letters Online and correspSearch, Epsilon uses Correspondence Metadata Interchange Format in TEI and open-source indexing tools to combine the work of large editorial projects, individual scholars, archive catalogues, and print correspondence sources to sift through the large volume of nineteenth century letters. The paper outlines our intentions to open a space that allows researchers to have a better sense of the rich variety scientific activities and actors during the period in which science as a professional discipline was being formed.

Digital Linguistic Landscape: the multilingual system of Ottawa

Herbert Natta
herbert.natta@gmail.com
Independent researcher

Maria Valese
m.valese@gmail.com
TU Delft

Keywords: linguistic landscape, data, map, minority
Topics: Europe, English, North America, Contemporary, social media analysis and methods, spatial & spatio-temporal analysis, modeling and visualization, Geography and geo-humanities, Linguistics

The proposal aims at investigating the digital linguistic landscape of Ottawa, analyzing social network geolocated data, to understand the structure of the

urban linguistic system (in terms of variety, use, typology), with a focus on minority languages.

Digital Public History: Origins of the Field and Possibilities for the Future

Sharon M. Leon
sharonmleon@gmail.com
Michigan State University, United States of America

Keywords: digital public history, historiography
Topics: English, North America, 20th Century, Contemporary, public humanities collaborations and methods, Galleries and museum studies, History

In recent years, award-winning scholarship from Denise Meringolo (Museums, Monuments, and National Parks, 2012) and Rob Townsend (History's Babel, 2015) have forced scholars to reconsider the historic relationship with between public history work and traditional academic history in the United States. That reconsideration has made historians return to the beginning of the 20th Century to account for a larger sense of the historical enterprise. Just as this work revises understandings of the larger discipline's history, digital humanities scholars are beginning question the long-accepted origin story of computational linguistics as the root of their field, including with some pointing to Public History (Robertson, 2014 and 2016). With this renewed attention to multiple origins, this paper will excavate the emergence of digital public history as distinct approach to collaborative engagement with the public, situating that emergence in the context of the history of public history and digital humanities.

A Collaborative Workspace for Archival Research: MIA and the EURONEWS Project

Lorenzo Allori
lorenzo.allori@gmail.com
University College Cork and Medici Archive Project

Carlotta Paltrinieri
carlotta.paltrinieri89@gmail.com
University College Cork and Medici Archive Project

Keywords: community-sourcing, self-digitization, self-archiving, preservation, collaborative research
Topics: Global, Europe, English, 15th-17th Century, 18th Century, crowdsourcing, digital archiving, History

A Collaborative Workspace for Archival Research: MIA and the Euronews Project

Introduction

MIA (Medici Interactive Archive - http://mia.medici.org) is a community-sourcing research portal, developed by the Medici Archive Project (http://www.medici.org), that uses self-digitization and self-archiving practices, along with textual data-entry, to reconstruct and preserve the Medici Archive held in the Florence State Archive. This archive comprising over 15 million documents dating from the 1370s to 1743 - later dispersed into ten collections - is one of the most extensive and most disrupted archives in Western Europe.

MIA allows a community of early-modern scholars to collaborate on research projects related to this archival corpus, and publish discoveries following a peer-review protocol. MIA is IIIF compliant, its backend is coded in JAVA (J2EE) and uses MySQL and MongoDB; its frontend is developed using AngularJS. MIA is open source and its code is available on BitBucket (https://bitbucket.org/map-firenze/mia and https://bitbucket.org/map-firenze/mia-fe).

Currently the portal contains more than 500.000 digitized documents (half of which on double folios), more than 25.000 transcriptions, 30.000 biographical entries, and 8.000 geographical and topographical records.

Private and Public Research

Physical accessibility is one of the main problems of archival research - as the latest Covid-19 pandemic has shown - and digitization offers a straightforward solution. However, conventional approaches to digitization are often laborious and costly. The self-digitization of primary sources presents a practical alternative. As this has become common practice in archives in Italy - especially after the Italian Ministry of Cultural Affairs allowed the use of smartphone photography in state archives - the need for a space in which to store, manage, catalog, and easily retrieve these images has become pressing.

In order to respond to the critical need for privacy during research, and the scholars' concerns over 'stolen

publications', MIA's uploaded material can be set as private or public, according to users' needs.

By opting for a community-sourced approach, and through a rigorous structure for data entering, MIA is able to maintain the integrity of the data associated with each document: this is a key advantage over crowd-sourced projects, which require a much greater effort. The beta-version was launched in late 2019: since then, over thirty scholars are using it in the spirit of open collaboration. To ensure greater scientific rigorousness, we are in the process of defining a vetting system to verify images and their metadata at various levels: from uploading images in the correct archival collection, to providing an accurate and standardized transcription, to identifying the people and places contained therein. MIA will be made available for the entire scholarly community in the Summer 2020, and the DH2020 conference will be the first occasion in which we present the entire platform with all its functionalities.

Peer-reviewed Publishing

A key aspect of MIA is the possibility of publishing important archival discoveries, in the form of Short Notices, directly on the portal - supporting the researcher's need for visibility and enriching their publication record - in an effort to elevate the publishing process within the humanities at a level of speed and efficiency comparable to that of the natural sciences. Once the user has filled out all the necessary metadata accompanying the image uploaded (date; description; typology; transcription; synopsis; people; places) they can click on the 'submit' button, and in a matter of (7-10) days will receive feedback from selected peer reviewers. Each Short Notice will be associated with an identifying number that will be used to cite the document in future scholarship.

Collaborative Research on and through MIA

MIA has been designed to support and encourage collaborative approaches to research. An extensive UX research task has been fulfilled in order to define the requirements for the development of the collaborative workspace. Moreover, we created a specific section in which the user can collaborate on topical projects with other MIA users: a space where to aggregate multiple documents and link them to other resources - such as PDFs, images, URIs, etc. For example, art historians can gather all the documents related to a specific artist (e.g. letters of commission and presentation, contracts, payment records, inventories) from otherwise separate collections in order to work collectively on them; while paleographers can cluster documents together in order to compare stylistic changes in handwriting or to identify an anonymous hand. The broad chronological, thematic, and typological range of the documents housed in MIA makes it a useful research tool for historians of art, medicine, music, religion, print culture, media, and more.

A fruitful result of a collaboration within MIA is the work done by the Irish Research Council EURONEWS project (University College Cork - http://euronewsproject.org), which focuses on analyzing pre modern newsletters in order to map their circulation and diffusion in early modern Europe. For this research project MIA provides a solid data-entry infrastructure, including: the uploading of self-digitized manuscripts; qualification of early modern documentary typologies with the appropriate metadata; additional XML encoding for the document transcription and synopsis fields to provide data interoperability for future integration with semantic models (RDF, Linked Data). This feature will support long-term availability and cultural analytics.

In addition to storing and classifying the newsletters, MIA has enabled Euronews Project to experiment with the data collected from all the documents written in the year 1600: this has led to further digital analyses that have brought to light issues and elements - such as unknown connections between specific cities and forgotten events that were the talk of European towns - that would have not been uncovered through a traditional approach or by an individual research project.

For those archival collections that have not been included in MIA, there is the possibility to create new branches within the git repository, or source code forks, allowing any library, archive, institute, or university to digitize, store, and share their material. One example is the newly created git branch for the Falconieri project (https://bitbucket.org/map-firenze/mia/branch/Falconieri). This project revolves around the study and the digital edition of a corpus of 430 letters sent to the Roman nobleman Ottavio Falconieri between 1655 and 1675. Thanks to MIA these letters, that have always been in the hands of private owners, will finally be accessible to everyone.

Conclusions

MIA successfully solves some of the most complex issues of archival research and traditional academic practices: digitally reconstructing the archive through community-sourced uploads; allowing for annotations, transcriptions, and organization of documents; ensuring the integrity of data; enabling collaborative work and discussion; and publishing archival discoveries in the form of Short Notices following a peer-review protocol.

Conceived from the start as a collaborative workspace, MIA supports non-linear research paths, gives space to interdisciplinary approaches, and encourages intersections of different disciplines and research methods. In its first public release, it offers

a rigorous and dynamic research environment for traditional historians and digital scholars alike.

Bibliography

Lewis Dafis L., Hughes L., and James R., 'What's Welsh for 'Crowdsourcing'? Citizen Science and Community Engagement at the National Library of Wales' in Crowdsourcing our cultural heritage edited by Ridge M., 153-156. London, Routledge, 2017.

Data-driven analysis of canonical works in early modern Britain

Mikko Tolonen
mikko.tolonen@helsinki.fi
University of Helsinki, Finland

Mark J. Hill
University of Helsinki, Finland

Ali Ijaz
University of Helsinki, Finland

Ville Vaara
University of Helsinki, Finland

Leo Lahti
University of Turku, Finland

Keywords: ESTC, canon, eighteenth century, public discourse, open science
Topics: Comparative (2 or more geographical areas)EuropeEnglishNorth America15th-17th Century , 18th Century, bibliographic analysis, spatial & spatio-temporal analysis, modeling and visualization, Book and print history, History

This paper studies how a work becomes canonical in early modern Britain. It is based on an extensive project on bibliographic records, and, the English Short-Title Catalogue (ESTC) in particular. We study different features of canonical works in detail. We aim to find relevant factors that canonical works have in common and examine the different paths works take in becoming a part of public consciousness.

Seven collections of prosody: the difficulty of visualizing non-hierarchical data

Rebecca Sutton Koeser
rebecca.s.koeser@princeton.edu
Princeton University, United States of America

Keywords: data visualization, prosody
Topics: Global, English, North America, Contemporary, Interface design, development, and analysis, Humanities computing

Why do so many data visualizations require sequential or hierarchical data? The sequential and hierarchical aspects of humanities data are often not the most important or interesting aspects. We've been exhorted to embrace the nonscalability and diversity of our data (Rawson and Muñoz), to disavow binary distinctions (D'Ignazio and Klein), and others are working to improve how we visualize uncertainty (Hullman). Creative and evocative visualizations abound, but they are rarely generalizable new approaches. Where are the new data visualizations for Digital Humanities work?

Based on work for a recent Princeton Prosody Archive (PPA) Editorial essay, I will share my experiences and failures attempting to visualize the seven scholar-curated collections within PPA. Categorization of the content in PPA is complicated because of the shifting meaning of the term prosody over time, encompassing both grammar and phonology as well as versification (Martin), and my visualizations reflect aspects of these two overlapping discourses in the form of the Linguistic and Literary collections within PPA.

The seemingly simple task of showing the relative size and overlap of all seven collections is actually quite difficult. A bubble plot or bar chart can convey relative sizes, but can't communicate overlaps and are limited to sets of similar scale. Visualizing the relative size and overlap of PPA (4,792 items) alongside the original bibliography that inspired its creation is useful, but put in context of all of HathiTrust (16 million items) we see a laughable diagram that nevertheless demonstrates the value of smaller, scholar-curated bibliographies (more than mere collections) based on HathiTrust materials.

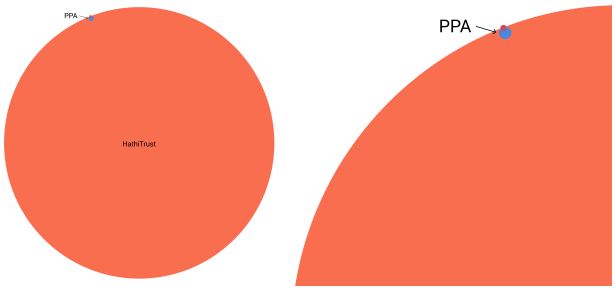


Figure 1. Venn diagram (left) and detail (right) of the relationship between PPA, the Original Bibliography and HathiTrust.

Relative sizes of bar charts are easier to read than circles, but a bar chart requires imposing an order. Euler and Venn diagrams (invented by mathematicians in the 1700s and 1800s) seem promising, but are limited in the number of sets, and also have the same problems with relative scale as bubble plots and bar charts. Treemap diagrams are useful for visualizing relationships, but only within a single hierarchy; likely because it was invented to visualize disk space utilization (Shneiderman and Plaisant). The UpSet plot is a relatively new solution to show "set intersections in a matrix layout" (Lex et al, 2014), which came out of bioinformatics, and is a powerful solution for this problem once viewers are oriented and trained to read an UpSet plot. Length proportional linear diagrams are another solution for set visualization (Rodgers et. al., 2015) which purport to improve set-related task performance, but they do not yet seem to be widely known or used.

After a quick overview of these visualization solutions and their limitations, I will demonstrate an experimental visualization I created as I worked on visualizing the PPA collections. It was inspired by the "warming stripes" climate change visualization (Hawkins) and a workshop series on p5.js called "Playing with Data" (Roth and Koeser). Vertical stripes are used to represent each item, with colored horizontal rows designating collection membership; interaction allows the viewer to focus on a single collection and see how it overlaps with others. In an earlier prototype, I sorted items somewhat arbitrarily by publication date. In my most recent prototype, items are ordered randomly to give viewers an overall sense of the distribution of the collections, but with interaction that allows a user to group all items in a single collection to allow focusing on overlaps with one collection at a time. Some might argue that a linear diagram would provide a more objective or efficient solution, but I'm less interested in viewers "accomplishing set-based task performance" than I am in the affective and aesthetic impacts (D'Ignazio and Klein) that make it possible to see and interact with these collections.

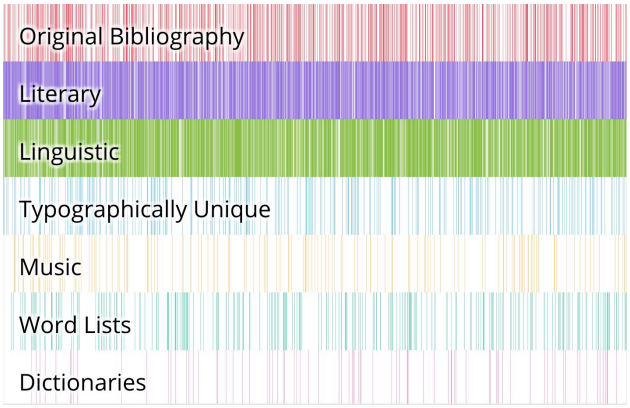


Figure 2. Screenshot of the prototype in its default view.

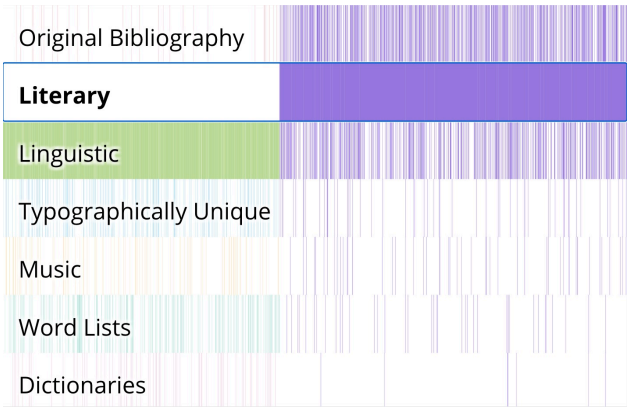


Figure 3. Screenshot of the prototype with the Literary collection selected.

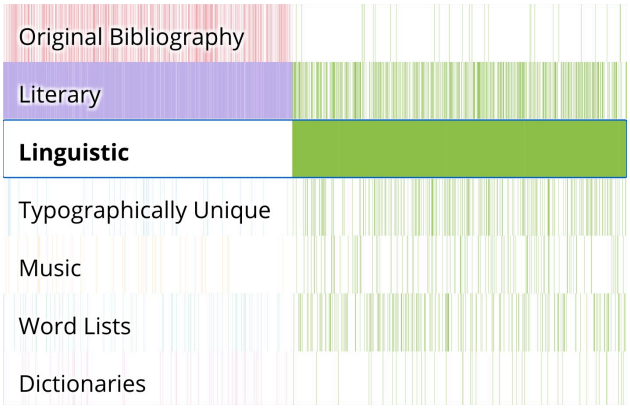


Figure 4. Screenshot of the prototype with the Linguistic collection selected.

Works Cited

Princeton Prosody Archive, version 3.2.4. Center for Digital Humanities at Princeton, 2018. http://prosody.princeton.edu. Accessed 04 October 2019.

D'Ignazio, Catherine, and Lauren F Klein. "Feminist Data Visualization." presented at the Workshop on Visualization for the Digital Humanities (VIS4DH), IEEE., Baltimore, 2016.

Hawkins, Ed. "Warming Stripes." Climate Lab Book (blog). Accessed October 11, 2019. https://www.climate-lab-book.ac.uk/2018/warming-stripes/.

Hullman, Jessica. "How to Get Better at Embracing Unknowns." Scientific American. Accessed October 18, 2019. https://doi.org/10.1038/scientificamerican0919-80.

Koeser, Rebecca Sutton. 2020. "Visualizing the Collections." Princeton Prosody Archive. January 17, 2020. https://prosody.princeton.edu/editorial/2020/01/visualizing-collections/.

Lex, A., N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. "UpSet: Visualization of Intersecting Sets." IEEE Transactions on Visualization and Computer Graphics 20, no. 12 (December 2014): 1983–92. https://doi.org/10.1109/TVCG.2014.2346248.

Martin, Meredith. "What Is Prosody?" Princeton Prosody Archive, version 3.2.4. Princeton: Center for Digital Humanities at Princeton. Accessed October 11, 2019. https://prosody.princeton.edu/prosody/.

Rawson, Katie, and Trevor Muñoz. "Against Cleaning." Curating Menus, July 6, 2016. http://www.curatingmenus.org/articles/against-cleaning/.

Rodgers, Peter, Gem Stapleton, and Peter Chapman. 2015. "Visualizing Sets with Linear Diagrams." ACM Transactions on Computer-Human Interaction 22 (6): 1–39. https://doi.org/10.1145/2810012.

Roth, Suzanne S. Roth and Rebecca Sutton Koeser. "CDH and CST Collaborate as They Play with Data!" Center for Digital Humanities at Princeton, March 18, 2019. https://cdh.princeton.edu/updates/2019/03/18/cdh-and-cst-collaborate-they-play-data/.

Shneiderman, Ben, and Catherine Plaisant. "Treemaps for Space-Constrained Visualization of Hierarchies," September 2014. http://www.cs.umd.edu/hcil/treemap-history/.

Aural Palimpsests: Voices, Space, History

Sanjin Muftic
sanjin.muftic@uct.ac.za
University of Cape Town Libraries, South Africa

Keywords: geofenced, performance, oral storytelling, audio walks
Topics: Africa, English, 19th Century, Contemporary, digital art production and analysis, spatial & spatio-temporal analysis, modeling and visualization, History, Performance Studies: Dance, Theatre

This oral presentation discusses how the emergence of public digital humanities has the potential to allow artists to reperform and re-investigate the histories of a place. It specifically looks at the development of new media technologies such as geo-fenced digital audio walks give listeners an opportunity to eavesdrop on a multitude of stories and voices, creating a palimpsest between the visible physical environment and the suggested future or imagined past.

Coming from a performance background, I look at how performative works can extend themselves into both a physical and virtual space through an investigation of history and public participation. I will present a tour of three such digital geo-fenced audio works staged in South Africa: "*ukubekindlebe*" - commemorating 200 years since the Battle of Grahamstown in today's Makhanda, "*All Who Listen*" - built out of extracts of the playscript All Who Pass capturing a family's last night in District Six, Cape Town prior to the forced removals of the 1970s, and a third yet unnamed project built upon public stories.

Using Achille Mbembe's claim that the "Internet is Afropolitan" (2015), this presentation makes a connection linking the oral histories of indigneous community storytellers to today's podcast culture. The circulation, migration, and transformation of stories across time and space, a feature of myths of origin, now is made manifest in our ears as we navigate a place. I examine how a critical appreciation can spur a reinvestigation and revoicing of history as a crucial part of our communities' futures. I further propose, within this area of Digital Humanities, libraries as the pointer towards these intersections, to archive, publish and share these aural works. So the library becomes a virtual space which is both on the internet and with the people who can tell you where to look and who to find. The librarians return to being the oral storytellers of the fourth industrial revolution.

References:

Bisschoff, L. 2017. The future is digital: an introduction to African digital arts, Critical African Studies, 9:3, 261-267, DOI: 10.1080/21681392.2017.1376506

Knowles, A. 2016. A more humane approach to Digital Scholarship. Items. Available: https://items.ssrc.org/parameters/a-more-humane-approach-to-digital-scholarship/

Mbembe, A., van der Haak, B. 2015. "The Internet Is Afropolitan." [online] The Chimurenga Chronic New Cartographies. March 17. Accessed August 12, 2015. http://chimurengachronic.co.za/the-internet-is-afropolitan/

Muftic, S. 2019. Ukubkeindlebe: How it Got Made. [online]. Available at: https://explore.echoes.xyz / collections/ZYrYXveaUPAk2SxY

Muftic, S., Jephta, A., Adams, Q., Landrew, E., Maarman, C., Petersen, J., Meyer, R., Isaacs, I. 2019. All Who Listen. [online] Available at: https://explore.echoes.xyz/collections/e0oWplw8aDb6t9AB

Muftic, S., Mbothwe, M., Makhetha, M., Rubela, A., Solontsi, Z., Maduna, L., Lebogang Mhlongo, K., Macanda, L, Lallie, L., Xakathugaga, N. 2019. Ukubekindlebe. [online] Available at: https://explore.echoes.xyz/collections/xwrvyQR5xMvTNBly

Munoz, T. 2012. Doing Digital humanities in the library isn't a service. Available: http://trevormunoz.com/notebook/2012/08/19/doing-dh-in-the-library.html. Accesses 25 August 2019.

O'Leary, N, Moreton, E. 2017. "The Migrant Letter Digitised: Visualising Metadata," Journal of Cultural Analytics. March 28, 2017. DOI: 10.31235/osf.io/89xbp

Pahl, M. 2016. Afropolitanism as critical consciousness: Chimamanda Ngozi Adichie's and Teju Cole's internet presence, Journal of African Cultural Studies, 28:1, 73-87, DOI: 10.1080/13696815.2015.1123143

Schreibman, S., Siemens, R., Unsworth, J. 2004. A Companion to Digital Humanities, ed. Oxford: Blackwell. Available: http://www.digitalhumanities.org/companion/

Short, H. 2010. Digital Humanities: Resources & Infrastructures CLARIN-NL Meeting University of Utrecht, 28 October 201. Available: http://www.clarin.nl/sites/default/files/utrecht-1010.pdf

Reutilización digital de datos públicos de la BNE como método heurístico para el análisis, mapeo y evaluación de las políticas culturales artísticas locales; aplicación a los sistemas expositivos temporales de Málaga y Valencia.

José Pino-Díaz
jpinod@uma.es
Universidad de Málaga, Spain

Nuria Rodríguez-Ortega

Universidad de Málaga, Spain

Keywords: catálogos de exposiciones, reutilización digital, datos públicos, evaluación de políticas públicas, política cultural
Topics: Europe, Spanish, Contemporary, cultural analytics, open access methods, Cultural studies, Political science

De entre el conjunto de actividades culturales que programan las administraciones públicas, como desarrollo de sus políticas culturales, las exposiciones temporales de arte constituyen uno de los eventos más importantes que suceden en las ciudades, tanto por su elevado coste para el erario público como por la gran cantidad de visitantes que acude a las mismas y por su trascendencia mediática.

"Hacer arte es hacer política", tal como afirma el artista Chema Lumbreras [1], y por extensión hacer exposiciones de arte es hacer política. Las administraciones públicas planifican exposiciones temporales de arte con fines políticos de propaganda o publicidad, de imagen, educativos o económicos [2].

Cultura y política local van de la mano en las exposiciones temporales de arte por su carácter de servicio público [3], siendo muy considerada su labor de divulgación del patrimonio cultural ante la sociedad. Las administraciones públicas además del objetivo social persiguen otros objetivos, García-Mestanza y García-Revilla (2016) apuntan que el desarrollo de los museos en la ciudad de Málaga se debe a una estrategia de fomento del turismo cultural [4].

Como en toda política pública se hace imprescindible la evaluación de las exposiciones temporales de arte, analizar si se han cumplido los objetivos cualitativos y cuantitativos señalados inicialmente, si los resultados obtenidos son los esperados o si el nivel de eficiencia en el empleo de medios y presupuesto es óptimo [5]. Una de las tareas fundamentales en la evaluación de políticas públicas es establecer un sistema de indicadores adecuado para la evaluación [6].

Los registros bibliográficos de la Biblioteca Nacional de España (BNE) [7] sobre los catálogos de exposiciones temporales de arte publicados en España proporcionan datos para el análisis cuantitativo, cualitativo y relacional y favorecen estudios epistemológicos sobre el arte como constructo social [8] y sobre las exposiciones temporales como las manifestaciones artísticas más lúdicas y populosas.

BNElab es el proyecto digital de la BNE [9] para apoyar la generación de nuevo conocimiento a partir de sus colecciones y datos. Una de las colecciones es la Bibliografía Española, que recoge los registros bibliográficos de los documentos publicados en España, recogidos por la BNE como depósito legal de los mismos.

Los nombres de los autores de los textos, artistas, comisarios o curadores, instituciones, entidad que publica y ciudad donde se publica, tal y como aparecen en el archivo XML exportado desde el BNE, asociados entre sí, mediante relaciones de co-autoría y co-ocurrencia, constituyen un sistema, es decir, un conjunto de elementos interrelacionados que contribuyen ordenadamente a construir y caracterizar el sistema de conocimiento "catálogos de exposiciones". Este sistema se puede mapear y visualizar, tomando las relaciones sistémicas de asociación (autores, artistas, curadores, términos del tesauro del título del tema, género/formato y entidad editorial), según el métod de análisis de similitud del software VOSviewer [10], metodología validada y empleada en estudios similares [11].

Se han estudiado los sistemas de exposiciones temporales de arte de las ciudades españolas de Málaga y Valencia durante el periodo 2017-2019 según los registros del catálogo de la BNE [12]. Los resultados indican que ambos sistemas de exposición se caracterizan por ser principalmente orientados y dirigidos por administraciones públicas y por una menor participación de las fundaciones o entidades privadas (ver figuras 1 y 2; redes de nodos con mínima ocurrencia igual a 1).

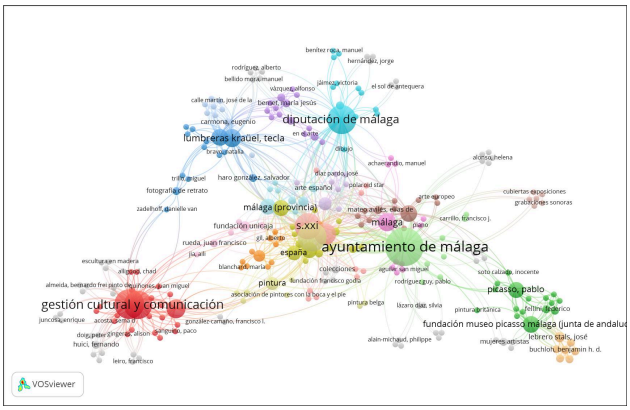


Figura 1: Sistema de exposiciones temporales de arte de la ciudad de Málaga (España) en el periodo 2017-2019. El tamaño de los nodos se ha hecho proporcional al número de catálogos de exposiciones de arte publicados e indexados en la BNE. Network visualization (nodos con mínima frecuencia 1).

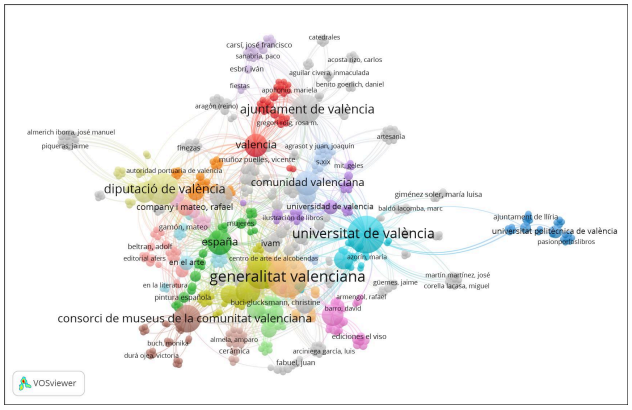


Figura 2: Sistema de exposiciones temporales de arte de la ciudad de Valencia (España) en el periodo 2017-2019. El tamaño de los nodos se ha hecho proporcional al número de catálogos de exposiciones de arte publicados e indexados en la BNE. Network visualization (nodos con mínima frecuencia 1).

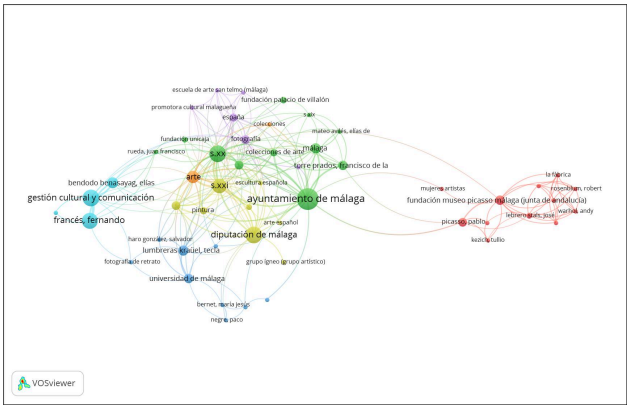


Figura 3: Sistema de exposiciones temporales de arte de la ciudad de Málaga (España) en el periodo 2017-2019. El tamaño de los nodos se ha hecho proporcional al número de catálogos de exposiciones de arte publicados e indexados en la BNE. Network visualization (nodos con mínima frecuencia 2)

En el sistema expositivo temporal de la ciudad de Málaga (ver figura 3; red de nodos con mínima frecuencia igual a 2) destaca el Ayuntamiento de Málaga por su posición central y por el número de catálogos (la Concejalía de Cultura es la responsable de la gestión del Museo de Patrimonio Municipal y una agencia municipal gestiona el Centro Pompidou Málaga, la Colección del Museo Ruso y la Fundación Picasso - Museo Casa Natal [13]) (ver figura 4; subred de nodos con mínima frecuencia igual a 2). Le siguen en importancia, aunque ocupando una posición periférica en el sistema expositivo, las siguientes instituciones: la Diputación de Málaga, la Universidad de Málaga, la Fundación Museo Picasso Málaga (entidad del gobierno regional de la Junta de Andalucía) [14 y 15] y la empresa Gestión Cultural y Comunicación [empresa privada que gestiona dos museos públicos mediante concesión administrativa, el Centro de Arte Contemporáneo (CAC) del Ayuntamiento de Málaga y el Museo de Arte de la Diputación Provincial (MAD)]. Con menor presencia y en la periferia de la red aparecen las instituciones privadas Fundación Palacio de Villalón y Fundación Unicaja y la entidad Red Promotora Malagueña.

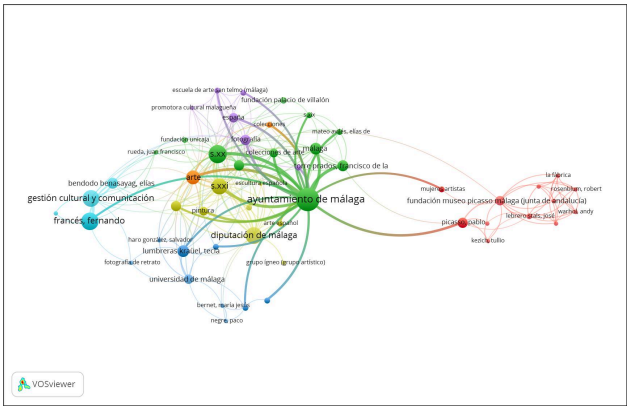


Figura 4: Sistema de exposiciones temporales de arte de la ciudad de Málaga; red del Ayuntamiento de Málaga (España) en el periodo 2017-2019. El tamaño de los nodos se ha hecho proporcional al número de catálogos de exposiciones publicados e indexados en la BNE. Network visualization (nodos con mínima frecuencia 2).

En el sistema expositivo temporal de la ciudad de Málaga se observan dos discursos expositivos principales: uno genérico, exposiciones de arte contemporáneo (s. XX y s. XXI) y de colecciones de arte, de pintura española y fotografía, sobre temáticas nacionales y locales; y otro específico, exposiciones sobre Pablo Picasso y su obra. Los artistas que aparecen en tres o más catálogos son Pablo Picasso (artista de referencia del sistema expositivo) y Francisco Jurado (artista local). Autores institucionales que aparecen en tres o más catálogos

son Fernando Francés (empresario de Gestión Cultural y Comunicación, director del CAC y del MAD), Elías Bendodo Banasayag (presidente de la Diputación Provincial de Málaga; autor institucional), Tecla Lumbreras Krauel (comisaria de arte y profesora de la Universidad de Málaga), Francisco de la Torre Prados (alcalde de Málaga; autor institucional) y José Lebrero Stals (director artístico del Museo Picasso Málaga).

En el sistema expositivo temporal de la ciudad de Valencia (ver figura 5; red de nodos con mínima frecuencia igual a 2) destacan también las iniciativas públicas, entre ellas, el gobierno regional, Generalitat Valenciana (ver figura 6; red de nodos con mínima frecuencia igual a 2). Son relevantes las subredes de las instituciones locales: Universidad de Valencia, Diputación de Valencia, Consorci de Museus de la Comunitat Valenciana y Ayuntamiento de Valencia. Como institución privada destaca la Fundación Bancaja, que se sitúa en la periferia de la red del sistema expositivo de la ciudad, igualmente a lo que sucede con las instituciones privadas en la ciudad de Málaga.

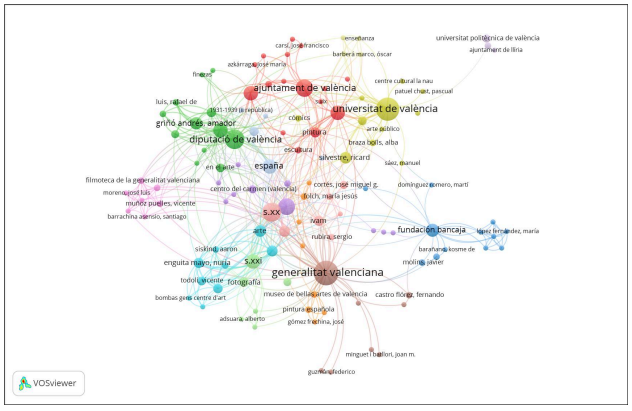


Figura 5: Sistema de exposiciones temporales de arte de la ciudad de Valencia (España) en el periodo 2017-2019. El tamaño de los nodos se ha hecho proporcional al número de catálogos de exposiciones publicados e indexados en la BNE. Network visualization (nodos con mínima frecuencia 2).

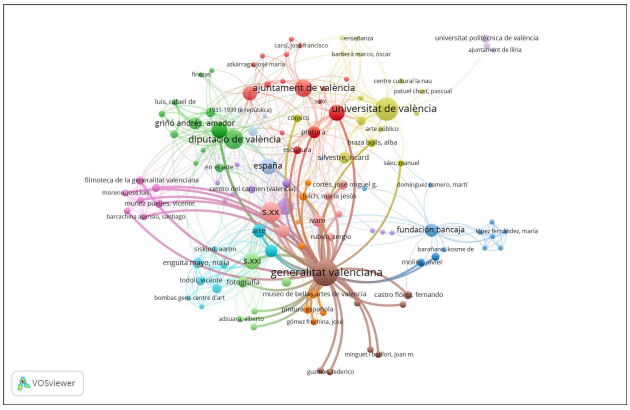


Figura 6: Sistema de exposiciones temporales de arte de la ciudad de Valencia; red de la Generalitat Valenciana (España) en el periodo 2017-2019. El tamaño de los nodos se ha hecho proporcional al número de catálogos de exposiciones publicados e indexados en la BNE. Network visualization (nodos con mínima frecuencia 2).

El sistema expositivo de la ciudad de Valencia se caracteriza por exposiciones temporales de arte contemporáneo (s. XX y s. XXI) y colecciones de arte de pintura española y fotografía, de temáticas nacionales, regionales y locales. También destacan las exposiciones indexadas con el término "historia de las exposiciones". Autores institucionales que aparecen en tres o más registros de catálogos son: José Miguel G. Cortés (director, comisario y autor de textos; IVAM), Vicente Todolí (director artístico y comisario; IVAM), María Jesús Folch (conservadora y comisaria; IVAM) y Sergio Rubira (subdirector de colección y comisario; IVAM); Rafael Company i Mateo (MuVIM); y Nieves López Menchero (jefa de documentación y publicaciones; Filmoteca de la Generalitat Valenciana). Comisarios que destacan son Amador Griñó Andrés (también figura como autor de textos), Joan Ramón Escrivá (también autor), Alba Baza Boils, Johanna Moreno Caplliure, Javier Molins y Laura Silvestre García (también artista). Como artistas destacan: Ricard Silvestre (también autor), Aaron Siskin y Rafael de Luis. Por último, como autores de los textos: Christine Buci-Glucksman, Fernando Castro Flores, Vicente Muños Puelles y Manolo Valdés).

Bibliografía

[1] N. Sánchez, «Hacer arte siempre es hacer política», eldiario.es. [En línea]. Disponible en: https://www.eldiario.es/andalucia/lacajanegra/Hacer-artesiemprees-hacer-politica_0_679632524.html. [Accedido: 20-oct-2019].

[2] A. Cohen, «¿Ritual o mangoneo de las exposiciones?», Ángulo muerto. [En línea]. Disponible

en: http://www.angulomuerto.com/index.php?cID=194. [Accedido: 20-oct-2019].

[3] B. Blasco Esquivias, «Las Exposiciones Temporales: Pasado, presente y futuro», Grupo Español GE-IIC. Curso sobre Exposiciones temporales y conservación del Patrimonio, 04-jul-2006. [En línea]. Disponible en: https://www.ge-iic.com/wp-content/uploads/2006/07/La_Exposicion_Temporal.pdf.

[4] J. García-Mestanza y R. García-Revilla, «El turismo cultural en Málaga. Una apuesta por los museos», Cultural tourism in Malaga. Museums challenge, jun. 2016.

[5] Instituto Andaluz de Administración Pública, «Evaluación de Políticas Públicas». [En línea]. Disponible en: http://www.juntadeandalucia.es/institutodeadministracionpublica/institutodeadministracionpublica/publico/evaluacion.epp [Accedido: 20-oct-2019].

[6] Instituto Andaluz de Administración Pública, «Guía para el análisis de la evaluabilidad previa de los planes de carácter estratégico». [En línea]. Disponible en: http://www.juntadeandalucia.es/institutodeadministracionpublica/institutodeadministracionpublica/publico/anexos/evaluacion/guiaevaluabilidad.pdf [Accedido: 20-oct-2019].

[7] Biblioteca Nacional de España, «Bibliografía española», 29-abr-2011. [En línea]. Disponible en: http://www.bne.es/es/Servicios/BibliografiaEspanola/. [Accedido: 20-oct-2019].

[8] J. R. Barbancho Rodríguez, «¿Puede el arte ser un constructo social?», Las liebres muertas, 10-nov-2010. [En línea]. Disponible en: http://juanramonbarbancho.blogspot.com/2010/11/puede-el-arte-ser-un-constructo-social.html. [Accedido: 20-oct-2019].

[9] Biblioteca Nacional de España, «BNElab. El proyecto de reutilización digital de la Biblioteca Nacional de España.», BNElab. [En línea]. Disponible en: https://bnelab.bne.es/. [Accedido: 20-oct-2019].

[10] N. J. van Eck y L. Waltman, «Software survey: VOSviewer, a computer program for bibliometri c mapping», Scientometrics, vol. 84, no. 2, pp. 523-538, ago. 2010.

[11] J. Pino-Díaz y D. Fiormonte, «La geopolítica de las humanidades digitales: un caso de estudio de DH2017 Montreal», DH2018, 26-jun-2018. [En línea]. Disponible en: https://dh2018.adho.org/la-geopolitica-de-las-humanidades-digitales-un-caso-de-estudio-de-dh2017-montreal/. [Accedido: 20-oct-2019].

[12] J. Pino-Díaz, «Aproximación epistemológica al sistema expositivo mediante la técnica de mapeo VOS, aplicación a los catálogos de exposiciones de la BNE publicados en Málaga y Valencia», RIUMA, 20-sep-2019. [En línea]. Disponible en: https://riuma.uma.es/xmlui/handle/10630/18415. [Accedido: 2-ene-2020].

[13] A. J. López, «El Pompidou, el Museo Ruso y la Casa Natal echan mano de la hucha», Diario Sur, 26-jul-2018. [En línea]. Disponible en: https://www.diariosur.es/culturas/pompidou-museo-ruso-20180726222130-nt.html. [Accedido: 20-oct-2019].

[14] A. J. López, «La Junta renueva su asignación al Museo Picasso Málaga», Diario Sur, 12-jun-2018. [En línea]. Disponible en: https://www.diariosur.es/culturas/junta-renueva-asignacion-20180612133013-nt.html. [Accedido: 20-oct-2019].

[15] A. J. López, «El Museo Picasso de Málaga luce sus cuentas», Diario Sur, 13-jun-2018. [En línea]. Disponible en: https://www.diariosur.es/culturas/museo-picasso-malaga-20180613230232-nt.html. [Accedido: 20-oct-2019].

(Re)collecting the Archive: Virtually Recovering Medieval Manuscripts Destroyed During World War II

Krista Alexandra Murchison
k.a.murchison@hum.leidenuniv.nl
Leiden University, Netherlands, The

Keywords: medieval manuscripts, book history, digital archives, XML, heritage preservation
Topics: Europe, English, 5th-14th Century, 20th Century, digital archiving, manuscripts description, representation, and analysis, Book and print history

Thousands of medieval manuscripts were destroyed during the Second World War. Among all the material losses of the war, these were especially tragic, since medieval manuscripts, by their very nature, are handmade and uniquely irreplaceable. Yet fragments of these lost artifacts remain in the form of editions, photographs, and descriptions. As this short paper will discuss, I am currently working on gathering together these fragments, thanks to funding from the

Netherlands Organisation for Scientific Research. The fragments will be used to digitally reconstruct the destroyed manuscript collections of four European nations and compiled into an online catalogue. Sections of the catalogue, where relevant, will be encoded in sustainable XML to protect these fragments against further loss. This catalogue will serve as a case study for exploring how digital technology, and the immaterial archival possibilities it entails, can help us think through the limits of the archive.

How Bisexual+ Community Members Tweet about Health: A Community-Informed, Topic Modeling Approach to Characterizing Under-Reported Health Needs of LGBTQ Twitter Users

Morgan Lundy
lundym@email.sc.edu
University of South Carolina

Frank Webb
University of South Carolina

Amir Karami
University of South Carolina

Keywords: Bisexual+ health needs, twitter, text mining, topic modelling, access to healthcare
Topics: English, North America, Contemporary, social media analysis and methods, text mining and analysis, Gender and sexuality studies, Library & information science

This study takes an intersectional and community-informed approach to characterizing and mining the Twitter accounts of thousands of users that self-identify openly in their Twitter bios as members of the Bisexual+ community. Twitter data was chosen due to the documented hesitations many Bi+ community members have about reporting their identities when completing official health documents. Accounts were assessed by multiple coders for inclusion in the dataset, in order to best represent the great diversity of the community while removing fake, automated and inactive accounts. The presenters share both the nuanced classification guidelines used for this project, and the process of developing guidelines that are responsive to community rhetoric and the intricacies of mining social media -- real, messy "data." The presenters also share on topic modelling methods and challenges, the nuances of natural language processing with tweets and the results found in the health topics Bi+ users share about online.

Inventing A Digital Chorapleth Map: An Alternative Choragraphic/ Psychogeographic Mapping Praxis

Clayton John Benjamin
cjbenjamin@ucdavis.edu
University of California, Davis, United States of America

Keywords: Psychogeography, Mapping, Choragraphy, Ulmer, Debord
Topics: English, North America, Contemporary, data modeling, spatial & spatio-temporal analysis, modeling and visualization, Cultural studies, Geography and geo-humanities

I present how I invented a digital mapping method/ genre I coin as a chorapleth map. Combining Ulmer's chorographic method and Debord's psychogeography method, I explain how a chorapleth map can be generated in juxtaposition to hegemonic practices embedded in producing choropleth maps. I explain how I applied the method for creating a chorapleth map to my consultation on the American addiction epidemic. Unlike choropleth maps, I was not interested in recording and relating amounts of overdoses in relation to addiction, I was interested in mapping the milieu and assemblage of addiction. I explain how I processed the data collected during my research on the American addiction epidemic and how I used that

data to create a chorapleth map. I argue that chorapleth maps are digital mapping genres appropriate for the digital age and can be used as a method within public digital humanities to advocate for alternative voices.

Physically-Based Rendering (Maya) for Experimental Modeling of Medieval Objects

Beth Fischer
blf1@williams.edu
Williams College, United States of America

Keywords: 3d visualization, art history, medieval, manuscripts, physically-based rendering
Topics: Europe, English, 5th-14th Century, manuscripts description, representation, and analysis, spatial & spatio-temporal analysis, modeling and visualization, Art history, Book and print history

This presentation demonstrates the value of the 3d graphics program Maya for modeling medieval manuscripts and other historic objects. Maya's physically-based rendering allows the creation of models that take into account the materials and finishes of objects. Objects can be animated or placed into environments that include varied lighting sources and other objects, allowing experimentation with different historical conditions of viewing.

The presentation includes a specific case study of this approach for a gilded early medieval gospel book, and also addresses the early stages of using Maya for modeling medieval textiles and liturgical objects.

Finally, I address the potential to share artifact models among different audiences and researchers, allowing users to combine and compare models within larger environments.

Sounding Spirit Overlays and Understories: OCR and Consortial Thematic Research Collections

Jesse P. Karlsberg
jesse.p.karlsberg@emory.edu
Emory University, United States of America

Keywords: thematic research collections, iiif, ocr, public digital humanities, American music
Topics: English, North America, 19th Century, 20th Century, Contemporary, digital libraries creation, management, and analysis, optical character recognition and handwriting recognition, Book and print history, Musicology

This presentation focuses on Sounding Spirit, an NEH-funded pilot public digital humanities project curating thematic research collections of digitized American vernacular sacred songbooks. Sounding Spirit uses Readux, a new platform developed by the Emory Center for Digital Scholarship that enables browsing, annotation, and publishing with federated collections of digitized books. First, I discuss how Sounding Spirit's corpus of nineteenth- and early twentieth-century songbooks poses unique challenges that require enhancements to contemporary optical character recognition (OCR). These challenges outline new avenues for research that will enhance textual scholarship with large corpora of digitized works. Second, I propose Sounding Spirit's model of interinstitutional collaboration as a replicable pathway for enhancing engagement with digitized books.

Mapping Topic Evolution Across Poetic Traditions

Petr Plecháč
plechac@ucl.cas.cz
Czech Academy of Sciences, Czech Republic

Thomas Nikolaus Haider
thomas.haider@ae.mpg.de
Max Planck Institute for Empirical Aesthetics, Frankfurt University of Stuttgart

Keywords: Topic Evolution, Poetic Tradition, Crosslingual
Topics: Comparative (2 or more geographical areas), Global, Europe, English, 18th Century, 19th Century, 20th Century, cultural analytics, text mining and analysis, Communication studies, Literary studies

Poetic traditions across languages evolved differently, but we find that certain semantic topics occur in several of them, albeit sometimes with temporal delay, or with diverging trajectories over time. We apply Latent Dirichlet Allocation (LDA) to poetry corpora of four languages: German (74k poems), English (85k poems), Russian (18k poems), and Czech (80k poems). We manually align and interpret salient topics, their trend over time (1600–1925 CE), showing similarities and disparities across poetic traditions with a few select topics, and use their trajectories over time to pinpoint specific literary periods with a focus on Romanticism and Modernism.

1 Corpora & Model

To determine the evolution of topics across poetic traditions, we collect four poetry corpora in Czech, Russian, German and English. See Table 1 for a size overview and where they were mined from. Our corpora cover a wide range of different genres and authors, but they are mildly contaminated with foreign language poems.

Language	Poems	Tokens	Comment
Czech	$\sim 80k$	15M	Corpus of Czech Verse (http://versologie.cz)
Russian	$\sim 18k$	2.7M	Poetic subcorpus of Russian National Corpus (http://ruscorpora.ru)
German	$\sim 74k$	12M	German Poetry Corpus v3, Textgrid + Deutsches Textarchiv (http://github.com/tnhaider/DLK)
English	$\sim 85k$	22M	Project Gutenberg, Mined with GutenTag 'Poetry' (https://gutentag.sdsu.edu/)

Table 1: Diachronic Poetry Corpora

To learn semantic topics, Latent Dirichlet Allocation (LDA) (Blei et al., 2003)has proved useful. We use the vanilla LDAMultiCore implementation as it is provided in genism (Rehurek and Sojka, 2011; https://radimrehurek.com/gensim/models/ldamulticore.html). LDA assumes that a particular document contains a mixture of a few salient topics of semantically related words.

We transform our documents to a bag of words representation and set the desired number of topics=100 and train for 100 epochs (passes) to attain a reasonable distinctness of topics. We choose 100 topics, as previous research on poetic topics (Haider, 2019; Navarro- Colorado, 2018)determined this parameter to be optimal for distant reading. Since we deal with highly inflected languages (Czech, Russian), lemmas were used instead of word forms. For lemmatization and POS-tagging of English and German texts we use the TreeTagger (Schmid, 1994),for lemmatization and POS-tagging of Czech texts we use the MorphoDita (Straková et al., 2014),for lemmatization of Russian texts we use MyStem (Segalovich, 2003). In Czech, German, and English all parts-of-speech (POS) except for nouns, adjectives, and verbs were filtered out. In Russian, the list of stopwords is provided by the NLTKlibrary and manually extended by us.

2 Experiment Setup

We approach diachronic variation in poetry as distant reading task to visualize the development of interpretable topics over time and across languages. We retrieve the most important (likely) words for all topics and interpret these (sorted) word lists as aggregated topics. We are then able to manually translate several topics that align over all four corpora.

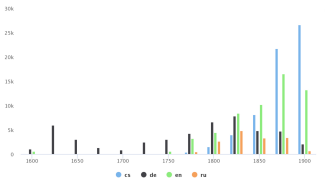


Figure 1: Size of Corpora over Time

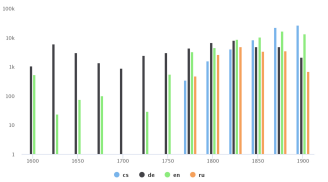


Figure 2: Size of Corpora over Time; log(Size) at y-axis

To discover trends over time, we bin our documents into time slots of 25 years width each, except for early English where two large slots (1600–1674 and 1675–1749) were used due to sparse data. See Figures 1 and 2 for a plot of the number of documents per bin. To visualize trends of singular topics over time, we follow the strategy of Haider (2019): We aggregate all documents d in slot s and sum the probabilities of topic t given d and divide by the number of all d in s. This gives us the average probability of a topic per time slot. We then plot the trajectories for each single topic.

3 Alignment and Interpretation of topic Trajectories

Based on a few selected topics, we can trace similarities and disparities over poetic traditions. See Figures 3–8 for a selection of interpretable topic trends where the four languages align.

Figure 3 shows the topic "Nation", which has a similar trend in German, Czech, and Russian, but is not present in the English corpus (cf. completely different geopolitical situation of the British empire). In the German corpus it emerges in the second half of the 18th century and peaks around 1825 to 1850 (outlining the period of 'Vormärz'). The same peak can be found in the Czech corpus (late National Revival), and slightly delayed in Russian. It loses importance in all three corpora after 1850/60, but it is gaining traction once again at the beginning of the 20th century.

Figure 4 shows the topic "Sea", which similarly rises towards the second half of the 19th century and

then stays stable into Modernism. This topic is most pronounced around the Russian and German period of Romanticism, after which it appears to taper off, while still following an upward trajectory into the 20th century for English and Czech.

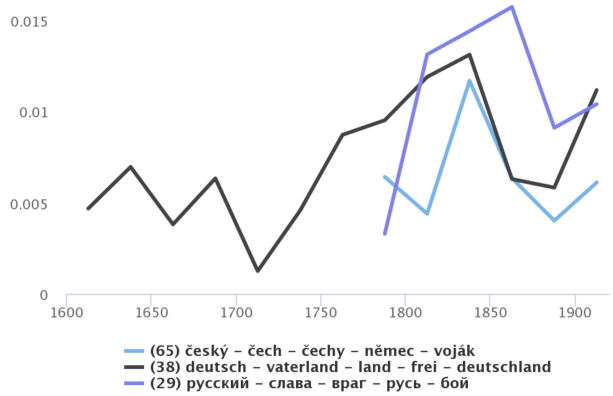


Figure 3: Topic **Nation**

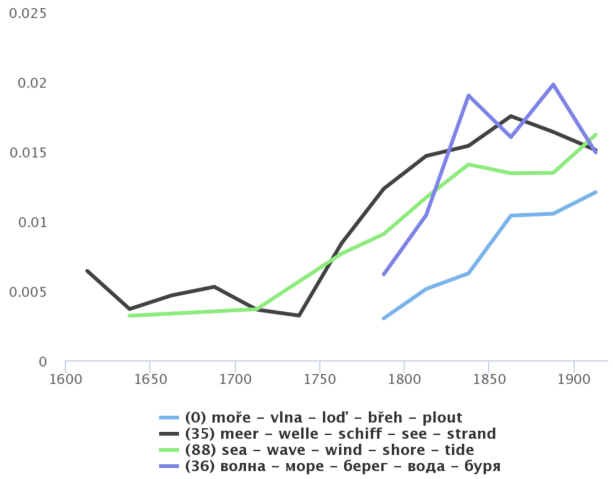


Figure 4: Topic **Sea**

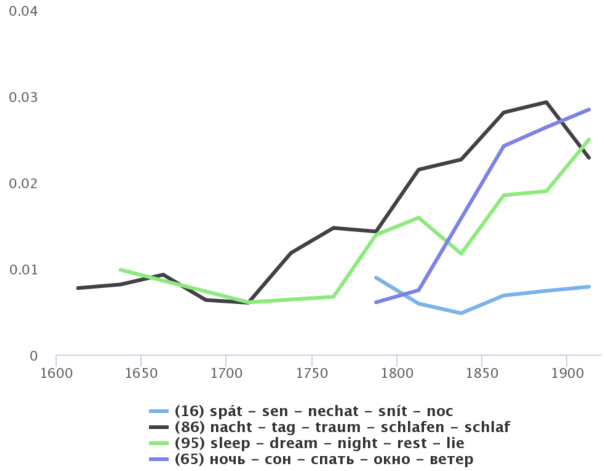


Figure 5: Topic **Sleep**

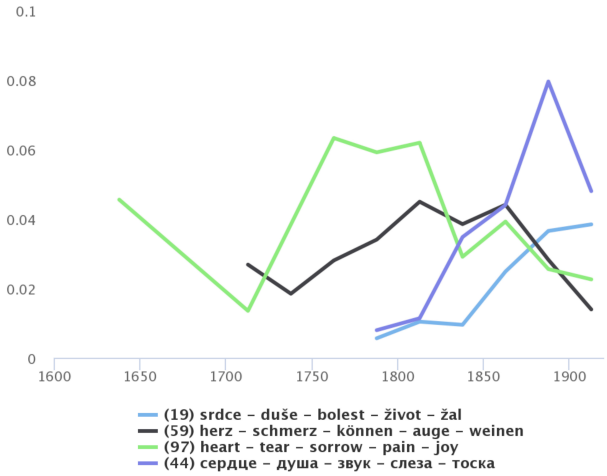


Figure 6: Topic **Sorrow**

The topic "Sleep" (Figure 5) appears quite correlated with the topic "Sea" in English, German, and Russian. While it is basically non-existent in the Early Modern Age (Baroque, Renaissance), it became increasingly popular toward late Romanticism and then Modernism, broadly delineating the long 19th century. Yet, it is rather marginal in the Czech corpus.

Figure 6 shows the topic "Sorrow", illustrating distinctly separate trends, with English and German on one hand and Czech and Russian on the other. For the first language pair it is associated with the period of Romanticism (although becoming prominent much earlier in English). Concerning the second language pair, we find this topic in late 19th century Modernism

(although in Russian it already emerges in the period of Romanticism, around 1825 to 1850).

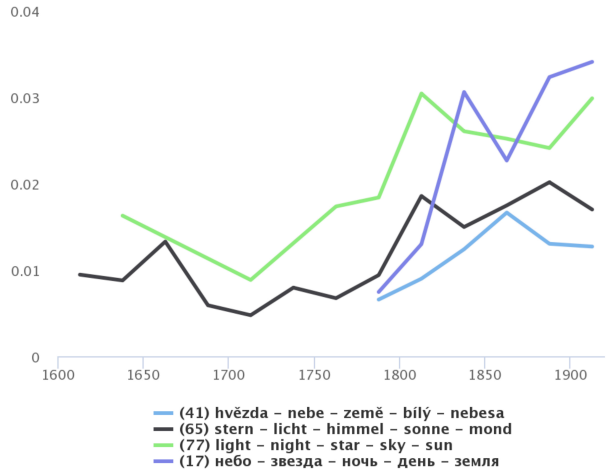


Figure 7: Topic **Stars**

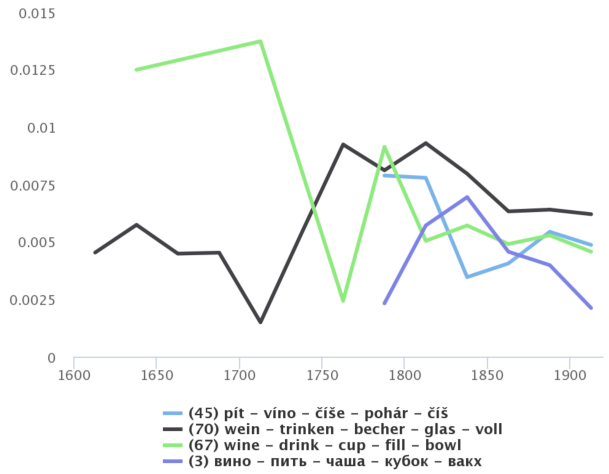


Figure 8: Topic **Wine**

Figure 7 shows the topic "Stars" that is pronounced in English and German High Romanticism (1800 to 1825) and in Russian Late Romanticism (1825 to 1850). In Czech the peak occurs delayed in the generation of "Máj" (period 1850 to 1875). Note that these authors claimed themselves as the followers of Karel Hynek Mácha (1810–1836), who in turn is well-known for bringing English Romanticism themes into Czech poetry.

Lastly, Figure 8 shows the topic "Wine", which is clearly associated with the Anacreontics. This topic is already quite visible before the onset of Romanticism, with accents in early 18th century English poetry, second half 18th century and later German poetry (High Romanticism), and late 18th century Czech poetry (almanacs edited by A. J. Puchmajer). In Russian poetry it surprisingly peaks in the period of romanticism (1825 to 1850).

4 Conclusion & Future Work

In this paper we used Latent Dirichlet Allocation for a visualization of topic trends across languages, illustrating the similarities and disparities between different poetic traditions. Our method is largely based on reading and translating topic distributions and finally interpreting the trajectories of relative topic importance against the backdrop of literary history. We find that some topics, especially the selected examples, do align across languages, sometimes with temporal delay (as they were picked up later in another language), while other topics were not as heavily discussed in other poetic discourses (such as "Nation" in English). In future work, we intend to look into cross-lingual alignment methods, e.g., through multi-lingual word embeddings or poly-lingual topic models without the need for parallel data. Finally, the over- or underrepresentation of certain authors or the presence of (near-)duplicates of poems (from different editions) can lead to corpus imbalance. Consequently, this impacts our measure to calculate the relative importance of a topic given a certain time stamp and should be addressed in future work.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Thomas N Haider. 2019. Diachronic topics in new high german poetry. Proceedings of the International Digital Humantities Conference DH2020 in Utrecht.

Borja Navarro-Colorado. 2018. On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry. Frontiers in Digital Humanities, 5:15.

Radim Rehurek and Petr Sojka. 2011. Gensim—statistical semantics in python. statistical semantics; gensim; Python; LDA; SVD.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In MLMTA.

Jana Straková, Milan Straka, and Jan Hajicˇ. 2014. Open-source tools for morphology, lemmatization, pos tagging and named entity recognition. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,

pages 13–18, Baltimore, Maryland, jun. Association for Computational Linguistics.

"Quantifying Kissinger" - Visualizing the Many Worlds of the Kissinger Correspondence

Micki Kaufman
mickikaufman@gmail.com
CUNY Graduate Center

Keywords: virtual reality, data visualization, computational analysis, foreign relations, archives
Topics: English, North America, 20th Century, public humanities collaborations and methods, virtual and augmented reality creation, systems, and analysis, History, Political science

This presentation will detail the latest from this Fortier Prize-winning doctoral dissertation, now nearing completion at the CUNY Graduate Center. The paper will present a brief history of the project's origins in foreign relations study, public history and archival research and detail its varying use of textual and network analysis, data visualization, narrative storytelling and, most recently, virtual and augmented reality.

Harry Potter and the Engaged Reader: Community Interactions and Influence in Serialized Fan Fiction

Mark Andrew Algee-Hewitt
malgeehe@stanford.edu
Stanford University, United States of America

Annika Butler-Wall
Stanford University, United States of America

Keywords: fan fiction, community engagement, text mining, reader feedback, social writing
Topics: Europe, English, North America, Australia/Oceania, Contemporary, cultural analytics, text mining and analysis, Literary studies, Sociology

Fiction does not exist in a vacuum: texts are the locus of a complex negotiation between authors, publishers, readers, and critics. This is particularly the case for serialized fiction: when authors publish a narrative over a series of installments, they open space for a feedback loop with readers. Engaged in an interactive community, invested readers become what Frank Kelleter calls "agents of narrative continuation."[1] The challenge for contemporary researchers is to reconstruct those communities in the absence of concrete material evidence.[2]

Current fan fiction communities can help us make sense of these complex relationships. Given their high publishing volume, as well as the opportunities for commenting, liking and following that the digital platform affords, fan fiction environments, offer unprecedented opportunities to study the relationship between authors and readers in serially published fiction.[3] Fan fiction communities value engagement: texts are often posted by chapter, and commented upon before the text is finished. As such, we can use these archives to explore the author/reader relationship.[4]

Our project asks two questions. First, what aspects of reader engagement can predict success for a text? And, in the context of specific stories, how are authors influenced by reader suggestions for content inclusion?

Corpus

For our corpus, we focus on the Harry Potter fan fiction community on fanfiction.net. By limiting our corpus, we control for different kinds of community-based interactions as well as subject and character. As our project is interested in serial publication of novel-length fiction, we acquired all Harry Potter fan fictions over 60,000 words in length, as well as all comments. Our final corpus consists of 16,454 texts with 1.8 billion words. Each chapter is associated with its full metadata, and with the text of its reviews.

Reader Feedback and Success

The fan fiction metadata offers quantitative measures of success to explore in relation to reader engagement. Readers mark stories as "favorites," providing positive feedback to authors and other readers. Using favorites as a measure of success, we discover that they are highly correlated to the total length of reviews written about the book (0.695). Furthermore, we divided all of the reviews, by chapter, into quartiles. Surprisingly, the most predictive reviews

of future success of the text are long reviews posted in response to chapters in the second quartile of the total text length (figure 1).

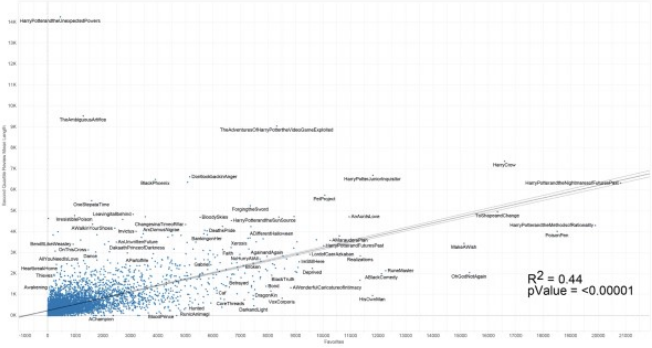


Figure 1: Regression of mean review length in second quartile on number of favorites

These findings indicate that the second quartile is perhaps the most important for securing reader interest (readers, unsure if the text will be completed, are likely less invested in the first quartile).

All of our results indicate a close reciprocal relationship between authors and readers in the fan fiction community. Moreover, by tracking the appearance of names in reviews that predate their appearance in the text, we argue that we have identified a proxy measurement for reader influence. In our full presentation, we not only show our results in far more detail, but we also explore the causal relationships between reader feedback and author response.

Reader Influence

Fan fiction is deeply invested in characters. A topic model run on the corpus returns a model wherein 78 out of 150 topics are predominately character names. Though fan fiction authors work with a fixed set of canonical characters, they make choices on which of those characters to include. We posit that when a character name appears first in the reviews, before it appears in the text, it may indicate that the author was prompted by the review, rather than the original novels. Restricting our model to character names that appeared more than two chapters later in the text (in order to account for false positives), more than 55% of stories contained at least one instance of a character name appearing first in the reviews, with an average 10-chapter delay between suggestion and uptake. Our data indicates that this is highly correlated with the success of the stories: texts in which the author appears to incorporate at least one character-name suggestion from a review has a higher probability of receiving more favorites (figure 2). Based on these results, the fan fiction community appears to be far more interactive than we had hypothesized: authors reliably achieve more success when overtly including suggestions from readers.

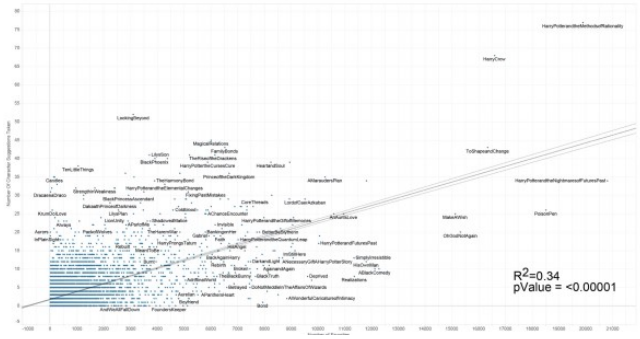


Figure 2: Regression of number of character suggestions taken on number of favorites

All of our results indicate a close reciprocal relationship between authors and readers in the fan fiction community. Moreover, by tracking the appearance of names in reviews that predate their appearance in the text, we argue that we have identified a proxy measurement for reader influence. In our full presentation, we not only show our results in far more detail, but we also explore the causal relationships between reader feedback and author response.

[1] Kelleter 2017, 13

[2] Studies in serial publication have largely taken a book historical approach, such as in Bonnell 2008, or Sigler 2008.

[3] Other DH studies have made use of fan fiction to research pedagogy (Black 2009), the expansion of Digital Humanities (Nicolaescu and Mihai 2014), and stylometry (Koolen 2018).

[4] See Milli and Bamman (2016) who also explore reader reactions to literature by predicting reader sentiment in response to text portions. Our interest diverges in that we are interested in the author's uptake of these reactions.

A tale of two web archives: Challenges of engaging web archival infrastructures for research

Jessica Ogden

jessica.ogden@soton.ac.uk
University of Southampton, United Kingdom

Emily Maemura
University of Toronto, Canada

Keywords: web archives, digital history, archival practice, digital source criticism, digital libraries
Topics: Europe, English, Contemporary, digital archiving, digital research infrastructures development and analysis, Humanities computing, Library & information science

Web archival collections are a key source for historical web research, and recent anthologies have provided examples of their use by researchers from a wide range of disciplines including the Digital Humanities. This paper explores the challenges of researcher engagement with web archives from the vantage point of two national web archives: the UK Web Archive at the British Library, and Netarkivet at the Royal Danish Library. We compare and contrast our experiences of undertaking web archival research at these institutions. Our personal interactions with the collections are supplemented by interviews and observations of the everyday practices of archivists, with a view towards understanding the similar yet different circumstances that shape researcher engagement in web archives. This analysis facilitates a discussion of the ways in which web archival interfaces both enable and foreclose on particular forms of research and contributes to critical ongoing debates surrounding digital source criticism in DH.

A digital assessment tool for monitoring and planning food security interventions in rural households of Uganda

Robert Tweheyo
rtweheyo@gmail.com
Kyambogo University, Uganda

Keywords: Digital assessment tool, food security, prevalence of food insecurity, intervention schemata, design sciences
Topics: Africa, South America, Comparative (2 or more geographical areas), English, 20th Century, Contemporary, digital ecologies and digital communities creation management and analysis, digital research infrastructures development and analysis, Humanities computing, Sociology

In this paper, a digital assessment tool that is used to monitor food security in rural homesteads is presented. It provides instant results in the form of statistics that can help in planning for appropriate interventions. Food security is an essential and universal measurement of community well-being. The fact that food security is a complex and multidimensional concept, its measurement has been a challenge to both researchers and practitioners. The tool makes food security assessment and monitoring easier and can generate instant statistics that can inform appropriate interventions aimed at enhancing decisions of rural communities on food security. The digital tool was evaluated with the intended users and evaluation results indicate that it is a useful device for monitoring food security and informing intervention programs.

Machine Learning for Literary Criticism: Analyzing Forms, Genres, and Figurative Language

Michael Ullyot
ullyot@ucalgary.ca
University of Calgary, Canada

Adam James Bradley
Adam.Bradley@uoit.ca
Ontario Tech University

Keywords: sonnets, poetics, genre, neural networks, topic modelling
Topics: Europe, English, North America, 5th-14th Century, 15th-17th Century, 19th Century, artificial intelligence and machine learning, text mining and analysis, Linguistics, Literary studies

Introduction / Importance

Since literary critics began using topic models on large text corpora, we perceive literary periods

as more fluid (Underwood, 2013; Pressman, 2014) and subgenres as more dynamic (Jockers, 2013; Underwood, 2014). These advances are mostly concentrated in prose fiction. Prose is more straightforward and verbose than poetry (Rhody, 2012), even if the problems of poetics are tractable for forms from Victorian sonnets to free verse (Houston, 2015; Bories et al., n.d.). Poetry is rarely straightforward: it uses words that resonate with other words, that complicate ideas and change meanings, that are there for idiomatic, rhythmic, allusive, formal, tonal, thematic, semantic, or idiosyncratic reasons. In sum, there are so many reasons that poets use particular words that machines struggle to model their topics statistically.

We use a recurrent neural network (RNN) for classifying sonnets, which are formally defined (14-line rhyming poems) but which also exhibit generic qualities of arguments, subjects/topics, tones, moods, and forms of address. We have built a computational model capable of scoring any text for its formal and generic resemblance to accepted criteria, for scoring its "sonnetness." Our goal is to find poems that have the generic features of sonnets, but not the formal criteria like a Petrarchan or Shakespearean rhyme scheme. These results will address our core question: to what degree sonnets, both individually and as a category, are defined formally or generically.

Methods

The standard distinction between Petrarchan and Shakespearean sonnets is based on rhyme schemes, but we set out to see if machine learning could define features that we couldn't see. We began with diction, or word choices that constitute both form and genre; the results were so promising that we extended the dimensionality of our model to incorporate four other dimensions: sound, rhyming, punctuation, and lineation. This identified a set of poems that we would never have considered.

Results / Discussion

In this presentation we will address why we began with early sonnets, which set conventions to which later English sonnets respond. We moved from a hand-transcribed test set to a corpus of 253,000 English-language poems from 12 centuries. Now we are expanding to two larger corpora: the 70,000 English texts printed before 1700, in the Early English Books Online - Text Creation Partnership (EEBO-TCP) corpus; and to the 334,000 volumes of literature in the HATHI Trust Digital Library.

In our past work (Ullyot and Bradley, 2018), we concluded that exceptions to the rules make language poetic. Poetry is deliberately irregular. It does not obey rules, it sets and then resets them. By expanding the canon of sonnets, our current project will unsettle critics' orthodox ideas about them.

Bibliography

Bories, Anne-Sophie, Petr Plecháč, and Pablo Ruiz Fabo (n.d.), 'Plotting Poetry / Machiner la poésie', https://machinerlapoesie.wordpress.com, Accessed 25 September 2019.

Houston, Natalie (2015), 'Visualizing the Cultural Field of Victorian Poetry', in Alfano, Veronica and Andrew M. Stauffer (eds.), Virtual Victorians: Networks, Connections, Technologies (New York: Palgrave Macmillan), 121-41.

Jockers, Matthew L. (2013), Macroanalysis: Digital Methods and Literary History, (Urbana, Chicago, and Springfield, IL: University of Illinois Press)

Pressman, Jessica (2014), Digital Modernism: Making it New in New Media, (New York: Oxford University Press).

Rhody, Lisa (2012), 'Topic Modeling and Figurative Language', Journal of Digital Humanities, 2 (1), n.p.

Ullyot, Michael and Adam J. Bradley (2018), 'Machines and Humans, Schemes and Tropes', Early Modern Literary Studies, 20 (2), n.p.

Underwood, Ted (2013), Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies, (Stanford: Stanford University Press).——— (2014), 'Understanding Genre in a Collection of a Million Volumes: Interim Report.' https://figshare.com/articles/Understanding_Genre_in_a_Collection_of_a_Million_Volumes_Interim_Report/1281251, Accessed 25 September 2019.

A TEI-XML Model for tone representation in African languages

Emmanuel NGUE UM
ngueum@gmail.com
University of Ngaoundéré, Higher Teacher Training College of Bertoua, Cameroon

Keywords: TEI-XML, tone, African languages representation, digital processing
Topics: Africa, English, Contemporary, annotation structures, systems, and methods, text encoding and markup language creation, deployment, and analysis, African and African American Studies, Linguistics

Language Documentation (LD) and description have generated textual resources for minority tone languages of Africa. With the event of computers, text resources are being created mostly in the form of digital-born texts. Visual layout of texts in these languages follows a variety of representation models, even within the same writing systems. Some of the issues at stake are IPA vs Latin characters; omission vs surface representation of tones as accents; surface vs deep representation of tones, one tier vs multiple tier representation of linguistic analysis.

This paper brings out the limitations of tone representation and analysis as an upper layer of character strings, and suggests a model based on the TEI-XML markup language.

Beyond Translation: language hacking and new pathways into language.

Gregory Crane
gregory.crane@tufts.edu
Tufts University, United States of America

Alison Babeu
Tufts University, United States of America

Elias Eells
Tufts University, United States of America

Maryam Foradi
Leipzig University, Germany

Camilla Rossini
University of Genova, Italy

Farnoosh Shamsian
Leipzig University, Germany

Sophia Sklaviadis
Tufts University, United States of America

Zach Sowerby
Tufts University, United States of America

James Tauber
Eldarion.com, United States of America

Allyn Waller
Tufts University, United States of America

Keywords: digital philology, reading
Topics: Comparative (2 or more geographical areas)Global, Europe, English, BCE-4th Century, 5th-14th Century, Contemporary, artificial intelligence and machine learning, scholarly editing and editions development, analysis, and methods, Language acquisition, Translation studies

Our work focuses on how audiences can push beyond translation, immediately engaging with the source text in the original language and internalizing as much knowledge of a language as their interest and available time warrant. The work-in-progress that we describe involves collaborators in Europe and North America, with funding from the US-based Mellon Foundation, the German DAAD, and the US National Endowment for the Humanities. Our work pragmatically exploits the rich body of openly licensed data available in Ancient Greek and the transnational position of Ancient Greek (along with languages such as Classical Chinese, Sanskrit, Classical Arabic, and Pre-modern Persian) but our work also engages with sources in languages such as Classical Arabic, Pre-modern Hebrew, Pre-modern Persian, and Early English and, in presenting at DH 2020, we hope to make contact with a wider range of languages.

How Do You Spell "😤"? The expansion of Unicode and the blurred line between text and image in digital space

S.E. Hackney
s.hackney@pitt.edu
University of Pittsburgh, United States of America

Keywords: Unicode, Emojis, Standards, Digital Text, FRBR

Topics: Global, English, North America, 20th Century, Contemporary, metadata standards, systems, and methods, text encoding and markup language creation, deployment, and analysis, Library & information science, Philology

This paper argues that while pictographs such as "😤" have been a part of Unicode since its initial launch in 1991, the 2010 introduction of emoji characters represents a major shift in the way that digital text is defined and standardized, and that this shift has major infrastructural and cultural implications for how we regulate and circulate language in digital space., It does this by examining the rules of the Unicode Standard, as well as through a reading of the official definitions and meanings-in-use of several emoji and non-emoji characters, and applying the LIS framework Functional Requirements for Bibliographic Records (FRBR) to how those characters exist and are circulated.

La obra de los científicos sociales del Siglo XX llevadas a las plataformas digitales del siglo XXI: los principios y prácticas de las humanidades digitales

Monica Calderon-Carranza
mcalder@pucp.pe
Pontificia Universidad Catolica del Peru, Peru

Enedina Ortega-Gutierrez
enedina.ortega@gmail.com
ITESM

Keywords: intelectuals, repositories, Google Scholar, Science 2.0, scholarly communication
Topics: South America, Spanish, 20th Century, Contemporary, digital archiving, open access methods, Library & information science

Dada la conmemoración del nacimiento del peruano Alberto Flores-Galindo fallecido en 1992, la Biblioteca de Ciencias Sociales de la Universidad Católica del Perú se planteó evidenciar su vigencia y sus escritos.

Se utilizaron prácticas, principios y valores de las humanidades digitales como el trabajo interdisciplinario, repositorios de acceso abierto, curaduría, metadatos, visualizadores de datos, gestores bibliográficos, indicadores bibliométricos y otras de la Ciencia 2.0. Se trabajó desde las prácticas y saberes propias de la bibliotecología, para dar a conocer la obra de AFG. Por otro lado, el trabajo interdisciplinario con docentes, permitió entender la dinámica de trabajo, la forma de comprender y analizar la obra intelectual desde la bibliotecología utilizando las tecnologías de acceso abierto. Conformamos un modelo para comprender la obra del "intelectual orgánico", llevado al repositorio y demostrando su actualidad usando Google Scholar Profiles, entre otros. Consideramos que este modelo puede replicarse en otros científicos sociales.

CollectionBuilder: Using Static Web Technologies to Create Digital Collections and Teach DH Principles

Olivia Wikle
omwikle@uidaho.edu
University of Idaho, United States of America

Evan Williamson
ewilliamson@uidaho.edu
University of Idaho, United States of America

Devin Becker
dbecker@uidaho.edu
University of Idaho, United States of America

Keywords: static web, Jekyll, digital collections, digital scholarship
Topics: Global, English, Contemporary, digital libraries creation, management, and analysis, physical & minimal computing, Education/ pedagogy, Library & information science

This presentation will discuss the development and deployment of the IMLS-sponsored project CollectionBuilder, an open-source tool that leverages

DH skill sets to help users create and host their own digital collection and exhibit websites via GitHub Pages. In addition to describing the tool itself, the presenters will demonstrate how faculty and staff from the University of Idaho's digital scholarship center, the Center for Digital Inquiry and Learning (CDIL), use CollectionBuilder to teach students fundamental DH skills and literacies through scaffolded learning embodied in the tool's customizable, modular design.

Stopifu: Supporting Task-Specific Stoplisting for Topic Models

Malcolm Mitchell
mitchellm@carleton.edu
Carleton College, United States of America

Eric Carlson Alexander
Carleton College, United States of America

Keywords: topic modeling, stopwords
Topics: English, North America, Contemporary, software development, systems, analysis and methods, text mining and analysis, Computer science

Probabilistic topic modeling is a promising and increasingly popular method of text analysis, affording the identification of patterns of change within tremendously large corpora of documents. However, though tools for exploring and analyzing topic models are increasingly common, the process of building a topic model remains something of an art, given the challenges of pre-processing and model training. To help make one stage of pre-processing more transparent, we have created Stopifu, a web-based tool designed to give researchers more direct control of stopword removal and help them anticipate the effect that excluding different words will have on their analysis. We present our design for this tool, along with our categorization of different types of stopwords that motivated its design.

Mapping Memory in Post-Hurricane Maria Puerto Rico By Combining Rhetoric and GIS

Gabriela Margarita Morales Medina
moral156@msu.edu
Michigan State University, United States of America

Ben Dougherty
dough101@msu.edu
Michigan State University, United States of America

Keywords: Rhetoric, GIS, Hurricane Maria, Uplifting Victims
Topics: South America, English, Contemporary, rhetorical analysis, spatial & spatio-temporal analysis, modeling and visualization, Cultural studies, Geography and geo-humanities

Academic intrigue has the potential to provide emancipatory power to disenfranchised communities by granting a voice to those who haven't been heard. In this project we aim to lift the voices of the people in Puerto Rico who were ignored after the passage of Hurricane Maria. We will do this by using qualitative methods from the field of Rhetoric and quantitative methods from the field of Digital Humanities to map out how people remember life during and post Maria on the island based on their descriptions and proximity to the event.

Interactive Transcripts and Multimodal Shakespeare

Olin Bjork
bjorko@uhd.edu
University of Houston-Downtown, United States of America

Keywords: Shakespeare, UDL, Drama, Multimedia Learning, Performance
Topics: Global, Europe, English, North America, 15th-17th Century, digital publishing projects, systems, and methods, Interface design, development, and

analysis, Disability and differently-abled studies, Education/ pedagogy

The interactive transcript is a multimodal presentation technique that became prevalent in online video platforms during the past decade and has recently begun to penetrate digital humanities. A transcript is interactive when it allows users to view the spoken dialogue within a video as a continuous/scrolling text synchronized with video playback (e.g., by means of a moving highlight) and to navigate to a desired location within the video by clicking on the corresponding section of the transcript (and vice versa). I am currently working with scholars at the University of Texas to develop a publicly available edition of Shakespeare's Julius Caesar that combines a video recording of a performance of the play with an interactive transcript of a text of the play. In my presentation, I will showcase this project to illustrate the practical, pedagogical, and scholarly rationales for interactive transcripts in multimodal editions featuring recorded dramatic performances.

Metrics for Distant Reading Literary History

Melanie Conroy
mrconroy@memphis.edu
University of Memphis, USA

Keywords: bibliometrics, literary history, movements, French literature, cultural analytics
Topics: Europe, English, 18th Century, 19th Century, bibliographic analysis, cultural analytics, Literary studies

The method of "distant reading," conceptualized most notably by Franco Moretti (2013), has been employed extensively in textual analysis. Analytics have most often been used for textual analysis in literary studies (Piper 2018). What should "distant reading" look like in prosopographical literary history? In this talk, I will examine various metrics that can be used to analyze and classify writers, as well as the potential significance of each metric. Using the concept of "sets," I classify writers as "globally successful," "nationally successful," "prolific," "minor," or "one-hit wonders," based on information on these writers' literary production derived from viaf.org data and combined with demographic data. The core dataset for this project is a database I have built of 10,000 French writers active at some time from 1700 to 1914 and a smaller subset of the top 1000 writers. This project uses the schema "Procope," developed as a part of the Mapping the Republic of Letters project at Stanford University (Comsa et al. 2016), combined with data from the National Library of France, to produce a large dataset of writers ranging from published authors of major works to correspondents of note, minor writers, and what I call "one-hit wonders," authors who have only one hit in library catalogues and are more likely to be unpublished authors or writers of correspondence than what we think of as canonical authors. The dataset also includes women writers from the Huygens database of women writers to address the lack of women writers in the other datasets. I use the following metrics from VIAF: the number of items held in libraries, number of current ISBNs, and the number of countries where materials are held.

In this new project, I consider various metrics which could be used to rank these writers, as well as similar projects by Ted Underwood and Mikko Tolonen with his group in Helsinki. I explore ways that writers could be grouped into sets and how these sets can be used to write alternative histories of literary and social movements by focusing on minor writers and figures who are generally marginal to histories of literary movements. The social and demographic data can be compared to "productivity" metrics from library catalogues, such as number of works published and number of editions, to make observations about which social networks (that is, academies, salons, movements) were more likely to produce successful or unsuccessful writers - both in the short term (in terms of contemporaneous publication) and in the long term (in terms of archiving and post-mortem publications). Finally, I look at which genres "failed" at a greater rate and the social status of "successful" and "failed" writers. Through this analysis, I produce a more nuanced picture of literary success than the division of writers into "canonical" and "non-canonical," especially by looking at writers who do not generally make it into literary histories at all.

Works Cited

Comsa, Maria and Melanie, Conroy Dan Edelstein, Chloe Summers Edmondson, and Claude Willan "The French Enlightenment Network," The Journal of Modern History 88:3 (2016), pp. 495-534.

Conroy, Melanie. Nineteenth-Century Networks: European Literary, Intellectual & Social Networks 1789-1914. http://blogs.memphis.edu/19thcenturynetworks/

Comsa, Maria and Melanie Conroy, Dan Edelstein, Chloe Edmondson, Claude Willan,

"French correspondents of major Enlightenment figures," Stanford Digital Repository (2014), http://purl.stanford.edu/bc436tm1194

Moretti, Franco. Distant Reading (London: Verso, 2013).

Piper, Andrew. Enumerations: Data and Literary Study (Chicago: University of Chicago Press, 2018).

Lahti, Leo, Marjanen, Jani, Roivainen, Hege, & Tolonen, Mikko. "Bibliographic Data Science and the History of the Book (c. 1500-1800)." Cataloging & Classification Quarterly, 57.1 (2019), 5-23.

Underwood, Ted. Distant horizons: digital evidence and literary change (Chicago: University of Chicago Press, 2019).

---. Why literary periods mattered: Historical contrast and the prestige of English Studies (Stanford: Stanford University Press, 2013).

Visualizing a Translational Queer Poetics

Corey Clawson

corey.clawson@rutgers.edu
Rutgers University - Newark, United States of America

Keywords: Archives, Graph Databases, Networks, Translation, Cultural Studies
Topics: Comparative (2 or more geographical areas), Global, English, 19th Century, 20th Century, network analysis and graphs theory and application, public humanities collaborations and methods, Gender and sexuality studies, Literary studies

Archivepelago is a project visualizing the transmission and translation of notions of sexuality and gender by mapping networks of queer writers and artists (and early sexologists), bringing into relief the communities that developed through these networks in the 19th and 20th centuries. The project draws upon hundreds of finding aids and additional biographic data, charting connections between these figures ranging from their correspondence to the works dedicated to and translated by one another. The project is intended to act as a public humanities resource for understanding the forces underpinning queer diaspora while encouraging scholars to rethink our conceptions of artistic influence beyond the misogynistic, heteronormative notions presented 35 years ago in Harold Bloom's The Anxiety of Influence.

In addition to drawing upon scholarly work on archipelagos (Manolo Guzman, Michelle Stephens, Island Studies Journal), this project draws upon Digital Humanities projects visualizing and recuperating historical networks such as Republic of Letters and Mapping Marronage as well as conversations about recovering and amplifying history via Digital Humanities projects.

Ultimately, Archivepelago's main features will be 1) interactive network maps demonstrating relationships in terms of correspondence, translations of one another's works, and shared demographic or psychographic characteristics such as religious affiliations; 2) interactive geographic maps depicting the migration of individual writers and artists, demonstrating emerging communities in metropolitan centers such as Paris, NYC, Algiers, and Mexico City; and 3) an online exhibit outlining key concepts drawing on well-known moments and figures (e.g., Walt Whitman, Gertrude Stein) and lesser-known figures like Charles Warren Stoddard.

In our current phase of the project, we are isolating translations of queer writers translating one another's work cataloged in finding aids and adding this information to our Neo4j graph database. We will share our findings by visualizing what Arjun Appadarai describes as "cultural flows" across languages (e.g., Langston Hughes' translations of Gabriela Mistral and Federico Garcia Lorca) and periods (e.g., John Addington Symonds' translations of Sappho and other Classical as well as Renaissance poetry) while putting our data in conversation with contemporary work on queer translations. Over the course of this presentation, we will illustrate the types of queries made possible by our graph database while discussing the significant figures and trends within these networks.

Caveat magister: a computational approach to designing a Latin curriculum

Daniel Libatique

dlibatiq@holycross.edu
College of the Holy Cross, United States of America

Dominic Machado

College of the Holy Cross, United States of America

David Neel Smith
College of the Holy Cross, United States of America

Keywords: Latin, morphology, pedagogy
Topics: Europe, English, BCE-4th Century, curricular and pedagogical development and analysis, natural language processing, Education/ pedagogy, Humanities computing

We introduce a code library for morphologically manipulating openly available corpora of Latin texts, and present a novel approach to designing a curriculum for learning Latin. We morphologically analyze a target corpus to sequence the presentation of grammatical features based on their frequency. We then identify sentence-level passages within the target corpus that include only a given set of features: in this way, for any stage in the curriculum, we automatically cull passages containing only forms and constructions that students have already learned. After four weeks in our fall, 2019, course, beginning Latin students were capable of reading unaltered continuous narratives that would have been impossible to approach with less than a semester and a half following a conventional textbook.

Human-Centered Computing for Humanists: Case Studies from the Computational Thinking and Learning Initiative at Vanderbilt University

Clifford B. Anderson
clifford.anderson@vanderbilt.edu
Vanderbilt University, United States of America

Corey E. Brady
corey.brady@vanderbilt.edu
Vanderbilt University, United States of America

Brian Broll
brian.broll@vanderbilt.edu
Vanderbilt University, United States of America

Lynn T. Ramey
lynn.ramey@Vanderbilt.Edu
Vanderbilt University, United States of America

Keywords: block-based programming, computational thinking, human-centered computing
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, text mining and analysis, Computer science, Education/ pedagogy

The Computational Thinking and Learning Initiative (CTLI) at Vanderbilt University formed out of an awareness that, across society and academe, computation is changing the nature of knowledge. As the practices and methods for producing, sharing, and contesting knowledge change, the enterprise of the university-its disciplinary scholarship, liberal arts mission, and charge to prepare professionals for the world of work-is being reshaped by algorithmic norms. The rise of computational thinking as a transdisciplinary category holds both promise and peril for the humanities. How can humanists, especially *digital humanists*, take advantage of the push for computational thinking across the curriculum while avoiding the dangers of appropriation? Alternatively, how can those developing and refining computational methods tap into the critical perspectives of the humanities? This paper discusses how an interdisciplinary group of colleagues is drawing on the theory of human-centered computing to develop environments and curricula for students of the humanities to explore the basics of text mining while also providing them with space to critique and resist the imposition of algorithmic rationality.

The CTLI foregrounds a particular image of computational thinking across the curriculum, working to identify, stabilize, and study new forms of human-computer partnership that are responsive to disciplinary ways of knowing. Our aim is to study new ways in which individuals or groups of humans, along with computers or groups of computational entities, can come together productively and critically as collective computational-thinking units and build on complementary strengths to investigate problems while rejecting facile technical solutions. The perspective of human-centered computing, that is, the "design of computing systems with a human focus from beginnin g to end,"[1] functions as an Archimedean point of our collaboration.

A trans-institutional group of researchers and scholars have assembled at the CTLI, including faculty from the schools of arts and science, engineering,

and education as well as the library and the data sciences institute. Together, we are exploring how human-centered computing collaborations transform epistemologies, practices, and pedagogies across disciplines.

Following the programmatic overview that Anderson and Ramey presented at DH2019,[2] these collaborators selected textual analysis as one of two focus areas (the other being climate change) for the first year of the initiative. Using NetsBlox,[3] a block-based programming environment developed at Vanderbilt University, we constructed components and curricula to teach students in the humanities the fundamentals of computational thinking (variables, looping, functions, recursion, etc.) by manipulating textual corpora rather than matrices of numbers. We wanted this environment to be "low threshold" enough for middle and high school students to use it as an entry point to disciplinary inquiry, while having a sufficiently "high ceiling" to allow scholars of literature and history to use it meaningfully as well.

Given our commitment to discipline-specific visions of computational thinking, our pilot project attempted to understand how the professional vision[4] of the humanist might resonate with the technological capacity of new computational tools and methods. We viewed human-computer collaboration as an integration of productively-different ways of interacting with the objects of analysis (e.g., texts). Computers and humans "read" differently, and the trick was to find a way to put these ways-of-reading into conversation. We used design-based research[5] to investigate possibilities, supporting and studying scholars' creative efforts to engage with technological tools and achieve what they regarded as progress on humanistic projects. We then developed a learning environment that enabled younger students to take on similar relations to computational tools, but in simpler, playful settings. Working with the same computational approaches at different levels of complexity and sophistication, we aimed to gain new perspectives on their power and limitations. We explored a design space with scholars and students centered on the analysis of style and affect in poetry, iteratively identifying and testing functionality with the ultimate objective of creating activities for both secondary-school and undergraduate courses.

In the course of our explorations, we narrowed the team's focus to three different technologies for textual analysis: (i) fundamental natural language processing concepts such as named entity recognition; (ii) word embeddings, such as Word2Vec;[6] and (iii) a "query runner" to TEI-encoded documents in BaseX,[7] a native XML database. During our collaboration, the team applied each of these technologies, first playfully, to explore what insights they could yield with familiar and constructed texts; then more deliberately, in settings in which it was likely that the computer's "readings" might be put in productive conversation with humans' readings. Finally, the team proposed questions, on the one hand, that revealed new patterns in larger corpora; and activities, on the other, that engaged younger learners in reflecting on style as a feature of writing under the writer's control.

In this paper, we report on this first year's effort and our progress in building a block-based computing environment and curriculum that supports textual analysis by both high school students and professors of literature. We discuss how these tools and perspectives scaffold activities as diverse as teaching secondary students about the linguistic differences between poetry and prose and detecting stylistic patterns among Victorian writers. We also discuss the limitations of taking a human-centered computing approach when conducting digital humanities research; in particular, we examine the drawbacks of block-based languages for text mining in comparison with tools like Lexos[8] and Voyant.[9]

Notes

[1] Jaimes, Alejandro, Nicu Sebe, and Daniel Gatica-Perez. "Human-centered computing: a multimedia perspective." In Proceedings of the 14th ACM international conference on Multimedia, pp. 855-864. ACM, 2006.

[2] Anderson, Clifford B., and Lynn T. Ramey. "Thinking Computationally in the Digital Humanities: toward block-based programming for humanists." DH 2019. Utrecht, 2019. https://dev.clariah.nl/files/dh2019/boa/0484.html.

[3] See https://netsblox.org/

[4] Goodwin, Charles. "Professional vision." American Anthropologist 96.3 (1994): 606-633.

[5] Cobb, Paul, et al. "Design experiments in educational research." Educational Researcher 32.1 (2003): 9- 13.

[6] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In Advances in neural information processing systems, pp. 3111-3119. 2013.

[7] See http://basex.org/

[8] See https://wheatoncollege.edu/academics/special-projects-initiatives/lexomics

[9] See https://voyant-tools.org/

Close-Reading the Endnotes for Word-Frequency Signposts & Polyvocal Citations

Sean A. Yeager
yeager.211@osu.edu
The Ohio State University, United States of America

Keywords: feminism, citations, bibliography, word-count, intersectional
Topics: English, North America, Contemporary, bibliographic analysis, meta-criticism (reflections on digital humanities and humanities computing), Feminist studies, Gender and sexuality studies

This paper describes a machine-assisted bibliographic analysis of the intertextual networks which are housed in an academic text's endnotes. My method is a three step process: first, run a word-count algorithm over the endnotes; next, identify patterns by subjectively grouping high-frequency words into loose clusters; finally, use these high-frequency signposts to perform a symptomatic close-reading of the endnotes. This project falls under Alison Booth's paradigm of "mid-range reading" because it sacrifices both the granular details of close reading and the massive scalability of distant algorithms. The payoff, however, is a versatile methodology through which digital research is consciously co-constructed with a mechanical collaborator.

My test case is Michael Hardt and Antonio Negri's Empire. I not only identify how monovocal scholarship can intertextually reify hegemonic hierarchies, but also how polyvocal citation can produce the opposite effect. I hope to prompt discussions about the intersections of digital scholarship.

The Potential and Danger of Digital History Pedagogy

Anelise Hanson Shrout
ashrout@bates.edu
Bates College, United States of America

Keywords: pedagogy, history, equity
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, network analysis and graphs theory and application, Education/ pedagogy, History

This presentation draws on a corpus of digital history syllabi from both the United States and Canada to explore the tensions between these varied uses and characterizations of digital history. It employs close readings of digital history syllabi alongside network analyses of co-assigned readings. The combination of these methods allows for a deep exploration the ways in which digital history is increasingly being used to challenge traditional power structures; to make space for students who are often marginalized within the academy; to share historical narratives beyond conventional audiences; while also deploying history in service of capital-enhancing activities and magnifying the voices of already dominant scholars; and dampening the voices of marginalized scholars.

Cherokee Syllabary Texts: Digital Philology and Experiential Learning

Jeffrey Gordon Bourns
j.bourns@northeastern.edu
Northeastern University, United States of America

Julia Hammond Flanders
Northeastern University, United States of America

Keywords: Cherokee, Philology, Linguistics, Digital Text Editions, Experiential Learning
Topics: English, North America, 19th Century, 20th Century, Contemporary, digital publishing projects, systems, and methods, manuscripts description, representation, and analysis, First nations and indigenous studies, Philology

The Digital Archive of American Indian Languages Preservation and Perseverance (DAILP) has partnered with the Linguistics program at project host Northeastern University in order to provide processing support for its digital text editions. Under this innovative arrangement, qualified Linguistics students provide philological processing and linguistic annotation of primary texts derived from digitized historical

Cherokee Syllabary manuscripts. The backdrop for these activities is a required research seminar that explores topics in Cherokee linguistics and provides hands-on philological training for working with Syllabary manuscripts and annotating these texts using the project's web-based collaborative software. The talk will describe this innovative approach to experiential learning and its vital contribution to language preservation and revitalization.

Inferring book relationships at the trillion-word scale

Peter Organisciak
peter.organisciak@du.edu
University of Denver, United States of America

Benjamin M. Schmidt
bs145@nyu.edu
New York University, United States of America

Keywords: text mining, large-scale digital libraries
Topics: Global, English, 20th Century, Contemporary, cultural analytics, text mining and analysis, Library & information science

Large digital libraries like the HathiTrust Digital Library (HTDL) provide texts of historical, cultural, or literary significance at unprecedented scales. However, the size and the consortial approach to building them can confuse computational attempts to model the collection, due to issues such as uneven duplication and incomplete metadata. This paper presents the technical workflow of a project seeking to address those challenges.

Can Digital Humanities Help in Finding Research Questions? A Comparative Analysis of the Attitudes Towards Neo-Confucianism Study of the Scholars Today and 300 Years Ago

Jun Wang
junwang@pku.edu.cn
Peking University, Beijing, People's Republic of China

Liyuan Zhang
zhangliyuan2013@pku.edu.cn
Peking University, Beijing, People's Republic of China

Haici Yang
Indiana University, Bloomington, United States of America

Keywords: Neo-Confucianism, Comparative Analysis, Network Analysis
Topics: Asia, English, 5th-14th Century, Contemporary, bibliographic analysis, network analysis and graphs theory and application, Asian studies, Philosophy

In this paper, we compare the results of our two studies: one takes Song Yuan Xue An, a book finished 300 years ago about the Neo-Confucianism history in Song and Yuan Dynasties, as the data source; while the other takes 5384 Neo-Confucianism-themed journal articles published over the past 30 years as the data source. In the first study, the overall development of Neo-Confucianism is divided into four phases through network analysis. But the numbers of contemporary research papers distributed over these four phases are significant unequal, thus indicating that the second phase, i.e. the period dominated by ancient scholars Yang Shi and Hu Anguo and their disciples, is neglected by contemporary scholars. These blind areas could lead to further humanities research questions.

Discovering Digital Humanities Methods Through Pedagogy A Study of Introductory DH Syllabi

Kristen Mapes

kmapes@msu.edu
Michigan State University, United States of America

Keywords: pedagogy, content analysis, training
Topics: Europe, English, North America, Contemporary, curricular and pedagogical development and analysis, meta-criticism (reflections on digital humanities and humanities computing), Education/ pedagogy

An understanding Digital Humanities as a field is enriched through exploration of the way that it is introduced to students. This presentation reports on a study that examined "Introduction to Digital Humanities" syllabi as a means of discovering the values and methods that shape the field and as a window into the intersection of the research and pedagogy practices of the instructors who teach introductory courses.

Constructing and Analyzing Short Science Fiction at Scale

Laure Thompson
laurejt@cs.cornell.edu
Cornell University, United States of America

David Mimno
Cornell University, United States of America

Keywords: science fiction, text segmentation, anthologies, corpus construction, text analysis
Topics: Europe, English, North America, 20th Century, Contemporary, text mining and analysis, Book and print history, Computer science

We evaluate methods for segmenting short stories from in-copyright science fiction anthologies that are only accessible through page-level word count data. Compared to the novel, short fiction is understudied in the digital humanities. This gap can be attributed to both the novel's role as the more popular (and prestigious) publication form in fiction and the added cost of separating collections into their constituent parts. While it is easier to study novels, novels are an insufficient proxy for studying the greater landscape of fiction. Many voices are lost if we limit our focus to novels. Moreover, the formation of science fiction as a distinct genre in English is directly tied to the rise of pulp magazines: a medium that focused on short fiction. Studying science fiction anthologies will allow us to find subtle trends in the history of the genre, and make quantitative comparisons between the structure of short stories and novels.

But accessing short fiction is difficult due to copyright, collection practices, and printing formats. Much of the work in science fiction falls into the mid-to-late 20th-century "dark zone" where it is too new to be public domain but too old to exist in born-digital format. The existence of large-scale repositories of digitized library collections, such as HathiTrust, offer the possibility to begin analyzing short fiction through non-expressive data views. The problem, then, is segmentation. The fundamental unit of HathiTrust is the volume, and little to no sub-volume structure is recorded in either HathiTrust metadata or the catalog records of source libraries. We therefore focus on identifying the boundaries of stories from non-expressive page-level data.

Data. We identify 690 volumes within the HathiTrust Research Center's Extracted Features Dataset (EF) (Capitanu et al. 2016) representing 529 science fiction anthologies published between 1940 and 2009. For each page in a volume, EF data segments headers, body text, and footers, and provides the text and part-of-speech of every word token in those segments, but does not preserve order within segments. For example, a header page of Nalo Hopkinson's "A Habit of Waste" within Women of Other Worlds: Excursions Through Science Fiction and Feminism is represented as [Hopkinson_NNP _JJ habit_NN waste_NN of_IN Nalo_NNP]. Note this data can include OCR errors, such as "\" instead of "A".

To validate our approach we also collect metadata from The Internet Speculative Fiction Database (ISFDB).[1] For anthologies, ISFDB typically provides the (edition-level) starting page for each short fiction entry, along with its title, author, length type (e.g. novella), and a list of other anthologies it has been published in, allowing us to link copies of the same story in different anthologies. We created a training set linking 68 EF volumes to their ISFDB metadata at the page level. This volume set represents 34 anthologies: two per anthology.

Segmentation. We treat the segmentation task as a supervised classification problem, in which the goal is to predict whether a given page is the beginning of a new segment. We evaluate a number of features for segmenting stories. For simplicity, we assume that every page corresponds to only one story. A reasonable starting point for identifying story transitions is to identify pages with unusually few tokens.

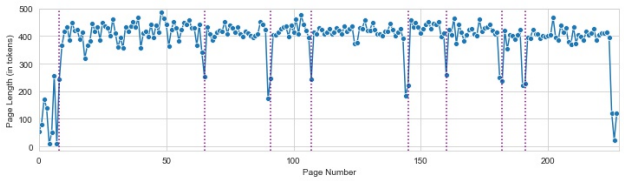


Figure 1: Page lengths can reflect actual transitions (vertical lines).

We note that there are often pairs of unusually short pages at segment boundaries, where one story ends mid-page and a new one begins in the middle of the next page. Since the absolute number of words per page varies from volume to volume, we normalize the length of each page to a number of standard deviations from the mean for the volume. For each page we also provide the deviation of the previous and following pages.

Headers and footers might also be helpful since they can contain short fiction titles and authors. We find that headers are much more common for science fiction anthologies than footers. There are 619 volumes with many headers, but only 20 with many footers. To allow for OCR errors, instead of exact string matches we measure the Jaccard similarity of character trigrams between headers. As headers often alternate, we record the maximum similarity of the five previous and subsequent pages and the maximum of the five subsequent pages. If boundary information is in the header, it becomes much easier to verify transitions given word statistics since these segments are very short.

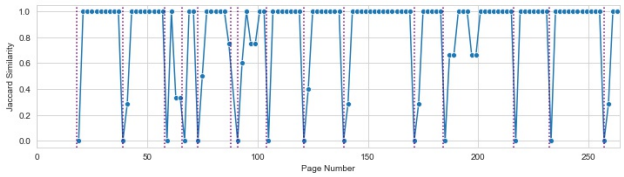


Figure 2: Similarity of odd page headers can also mark transitions, but sometimes they're absent.

Analysis. In addition to raw features from page length and header similarity, we add a small set of pairwise product features to represent interactions between variables: length * prev_length, length * next_length, and the product of those two products; prev_header_similarity * next_header_similarity. For analysis, we use 34 labeled volumes representing distinct anthologies. For 20 cross-validation splits we hold out 20% of these volumes for evaluation. Using all features, we find that 58% +/- 11.8 of labeled boundaries match a predicted boundary. 63% +/- 10.9 of labeled boundaries have a predicted boundary within one page. Boundary predictions are conservative, with 93% +/- 3.7 of predictions within one page of a labeled boundary. The difficulty of the task varies between volumes, often because of the quality of headers, OCR problems, and idiosyncratic boundaries. Without product features, performance drops to 37% exact, 44% of labels within one page of a prediction, and 79% of predictions within one page of a label. With only raw length features these numbers are 14%, 17%, and 67%; header features alone are precise but limited, at 21%, 26%, and 78%. Results for the other 34 labeled volumes are similar.

Applying the full model trained with all labeled volumes to the complete set of 690 anthologies results in just over 10,000 predicted segments. Not all of these correspond to stories, but even with this noisy sample we have substantially improved our ability to access short-form science fiction.

Future Work. All of our current metrics rely on the formatting of pages. We predict that the content of stories may also have clues, with the names of characters and settings changing between stories. Although we expect these methods to be less successful, they will be useful indicators for when an identified segment needs to be further divided and will work for a wider range of publishing formats. The fact that product features did so well indicates that structured prediction models for sequences and multi-layer neural networks might improve performance further.

Notes. 1. https://www.isfdb.org

References

Capitanu, Boris, Ted Underwood, Peter Organisciak, Timothy Cole, Maria Janina Sarol, and J. Stephen Downie (2016). The HathiTrust Research Center Extracted Feature Dataset (1.0) [Dataset]. HathiTrust Research Center, http://dx.doi.org/10.13012/J8X63JT3.

Archives, Podcasts, and Place-Based Audio Storytelling

Jim McGrath

james_mcgrath@brown.edu
Brown University, United States of America

Keywords: podcasts, digital storytelling, audio storytelling, sound studies
Topics: English, North America, 20th Century, Contemporary, digital archiving, public humanities

collaborations and methods, Cultural studies, Media studies

This short presentation surveys podcasts and other forms of audio storytelling that remediate or otherwise make extensive use of archival materials from particular collections and institutions. It then focuses on recent efforts in a specific public humanities academic context to create podcasts and place-based audio storytelling projects that build on similar efforts from journalists, comedians, and cultural institutions. In distinguishing between "podcasts' and "place-based audio storytelling," this presentation also hopes to demonstrate how relationships between audience and place can be created and critiqued in this digital storytelling format.

A Digital Study of Ralph Ellison's Integrative Form

J.D. Porter
jdporter@stanford.edu
Stanford University, United States of America

Keywords: Ellison, Classification, Grammer, Influence, Style
Topics: English, North America, 20th Century, text mining and analysis, African and African American Studies, Literary studies

In this project, I use classification of textual segments on the basis of parts of speech to examine how Ralph Ellison processes the form of his literary influences into the text of his canonical novel Invisible Man. The project is both a proof of concept for the use of DH in Ellison scholarship, where digital approaches remain quite rare, and a demonstration that parts of speech can successfully indicate authorial style and influence.

Digital History, Deep Mapping and Public Heritage in the Undergraduate Classroom

Austin Mason
amason@carleton.edu
Carleton College, United States of America

Keywords: Deep Mapping, GIS, Pedagogy
Topics: English, North America, 19th Century, 20th Century, Contemporary, public humanities collaborations and methods, spatial & spatio-temporal analysis, modeling and visualization, Geography and geo-humanities, History

Can new methods of digital humanities and collaborative public history foster an appreciation of the diversity of experiences of space and place through time and across cultures? This paper presents a case study of a recent undergraduate seminar that sought answers to this question through a deep mapping of the long history of Northfield, MN, before and after its most well-known era of late nineteenth-century European settlement. Students were introduced to digital mapping technologies including Geographic Information Systems (GIS), web mapping frameworks, 3D modeling, and video game engines, which they used to map the long history of the region from geological time through its historical origins as part of the Dakota homelands, to its recent past and present as a region of increasingly diverse immigrants. The results (while still a work in progress) prove that digital deep mapping methods are particularly well suited for sharing human-centered stories grounded in land, region and community.

Deep mapping is as much archaeology as it is cartography, plumbing the depths of a particular place to explore its diversity through time.[1] The term was coined by William Least Heat--Moon, whose book *PrairyErth* set the standard as a richly layered, historical and literary study of the land and peoples of Chase County, Kansas.[2] "Deep mapping" has since been taken up by the digital and spatial humanities communities to explore how issues like time depth, multivocality, and uncertainty — notoriously hard to represent in traditional printed maps — can be visualized using today's digital mapping and interactive programming techniques.[3]

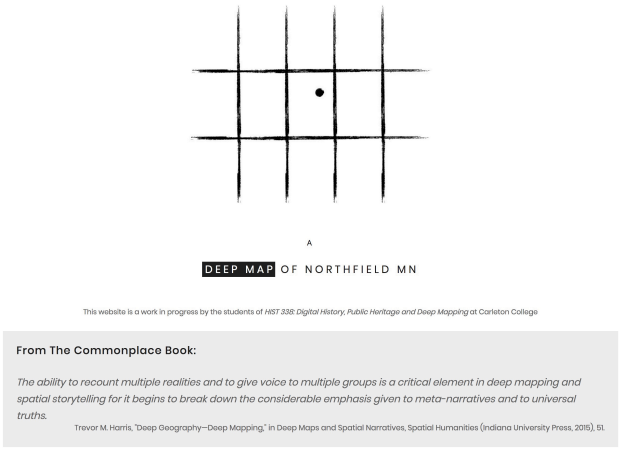


Figure 1: "A Deep Map of Northfield MN," a digital remix of PrairyErth's main elements

Over the course of 10 weeks, the students and I collaboratively built a "deep map" website that tells a series of spatial stories, built in numerous platforms and organized geographically, but linked together through modern web development tools.[4] Inspired by *PrairyErth*, our goal was to construct a deep map that "encompasses a range of approaches—cartographic, geographical, historical, literary, philosophical, scientific, anthropological, sociological and theological—to weave multiple strands of evidence, observation, impression and memory into a distillation of place."[5] Least Heat-Moon organized his book (somewhat arbitrarily) by US Geological Survey quadrangles, further dividing those sections into chapters that focus on one physical or thematic aspect of the territory. Our website would do the same, with sections devoted to each quadrangle surrounding Northfield MN and 3-5 sub-pages narrating and illustrating distinct aspects of that quadrangle's spatial history.

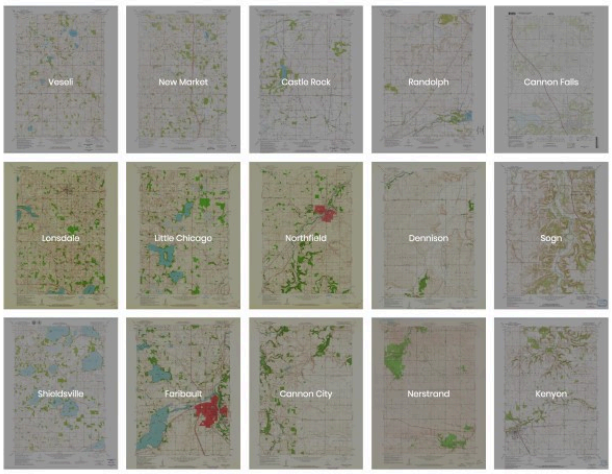


Figure 2: Primary navigation of the website is via USGS Quadrangle maps

The key innovations of our remix of PrairyErth's main elements were: 1) multivocality, allowing numerous authors to deepen each other's insights, and 2) leveraging digital mapping technologies to make these "chapters" interactive and publicly accessible through the internet. ESRI StoryMaps narrate the fraught migrations in the region, from ancestral Dakota peoples, to Czech and Norwegian immigrants, to the current increase in Somali and other West African populations. 3D GIS scenes and embedded Unity games allow surrogate experiences of movement through the landscape, while custom scripts juxtapose individual maps with quotations from our collective Common Place Book (a concept again borrowed and digitally updated from Least Heat-Moon).

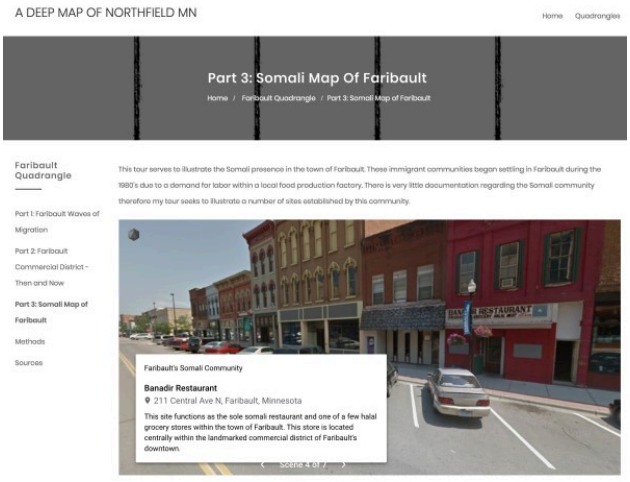


Figure 3: "Somali Map of Faribault," by Zoe Denckla '20 using Google Tour Creator

The aim of this paper is to share our experiences and open a dialogue about deep mapping as a method for collaborative public scholarship that can connect engaged undergraduate learning with the needs of community partners while staying firmly grounded in place.[6]

[1] Les Roberts, "Deep Mapping and Spatial Anthropology," Humanities 5, no. 1 (January 14, 2016): 5, https://doi.org/10.3390/h5010005.

[2] William Least Heat-Moon, PrairyErth: A Deep Map, Reissue edition (Boston: Mariner Books, 1999).

[3] David Bodenhamer, John Corrigan, and Trevor Harris, eds., Deep Maps and Spatial Narratives, Spatial Humanities (Bloomington: Indiana University Press, 2015).

[4] "A Deep Map of Northfield MN," Austin Mason et al. 2019, https://medhieval.com/classes/deepmapping/

[5] Carenza Lewis, "Archaeological Excavation and Deep Mapping in Historic Rural Communities," Humanities 4, no. 3 (September 10, 2015): 393–417, https://doi.org/10.3390/h4030393.

[6] Denis Wood, "Mapping Deeply," Humanities 4, no. 3 (August 6, 2015): 304–18, https://doi.org/10.3390/h4030304.

Implementation of Event Classification and Spatio-temporal Information Integration Analysis based on Meteorological Records in Historical Texts: A Digital Humanities Case Study on Climate Change and Social Trends

Shi-Yun Huang
Department of Computer Science and Information Engineering, National Central University, Taiwan

Shang-Yun Wu
Department of Computer Science and Information Engineering, National Central University, Taiwan

Yu-Chun Wang
Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts, Taiwan

Cheng-Han Wu
Department of Computer Science and Information Engineering, National Central University, Taiwan

Jung-Yi Tsai
thtsai@csie.ncu.edu.tw
Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

Richard Tzong-Han Tsai
Department of Computer Science and Information Engineering, National Central University, Taiwan
Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan

I-Chun Fan
Institute of History and Philology, Academia Sinica, Taiwan

Keywords: Climate Event, BERT, Spatio-Temporal
Topics: Asia, English, 15th-17th Century, 18th Century, natural language processing, spatial & spatio-temporal analysis, modeling and visualization, History, Environmental, ocean, and waterway studies

Seeing the vision of things is the great application concept of Science Research. Since extreme climate event occurs more often than before, the importance of the research about the historical climate texts increases as well. Tracing back to the contemporary occurrence and the impact of climate disasters, we could find the cues in a rich database from early Chinese historical texts. By extracting the natural phenomena in the historical texts, analyzing the environmental possibilities and restrictions, learning on the impact on social phenomena at the time, we gain insight into the causes and effects of social pulsations. This study takes the 'China's Three Thousand Years of Meteorological Record' (2013) which collects more than 7,000 historical records of as the text source, categorizing from regions, dates, presenting the visualize data on the graphic platform and analyze the spatial and temporal characteristics of social humanity events caused by nature phenomena.

An Advising Framework for Intersecting Disciplines within Graduate Digital Humanities Certificate Programs

Rennie Mapp
mapp@virginia.edu
U of Virginia, United States of America

Keywords: advising, DH Certificate
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, Education/ pedagogy

This short paper describes an advising framework developed for a graduate digital humanities certificate program recently established at an R1 university.

A Visualization-Assisted Reading Systemfor a Neo-Confucian Canon

Jun Wang
junwang@pku.edu.cn
KVision Lab, Department of Information Management, Peking University

Xiaoyu Li
KVision Lab, Department of Information Management, Peking University

Enhua Bian
KVision Lab, Department of Information Management, Peking University

Linxu Wang
KVision Lab, Department of Information Management, Peking University

Shuran Liu
KVision Lab, Department of Information Management, Peking University

KVision Lab, Department of Information Management, Peking University

Nuo Chen
KVision Lab, Department of Information Management, Peking University

Keywords: visualization, Song-Yuan Xue'an, knowledge graph, Chinese classic literature
Topics: Asia, English, 5th-14th Century, linked (open) data, spatial & spatio-temporal analysis, modeling and visualization, Asian studies, Library & information science

Abstract This paper aims to reconstruct *Song- Yuan Xue'an* into a knowledge graph and design a visualization-assisted and interactive reading system for it. It seeks to make this Chinese classical writing more understandable for the younger generation, and meanwhile to provide a useful indexlike instrument for humanities researchers. The paper presents two steps of reconstructing *SongYuan Xue'an.* First, the text is converted into a knowledge graph by extracting named entities of scholars, places, dates and their inter - relationships, to build a graph database with Neo4j platform. Second, three visualization tools are used to demonstrate the genealogy of philosophy schools and relationships among them.

Keyword visualization; *Song-Yuan Xue'an;* knowledge graph; Chinese classic literature

1 INTRODUCTION

Song-Yuan Xue'an (Scholarly Annals of Song and Yuan Dynasties) is a historical account of philosophers and their doctrines in Song and Yuan Dynasties of ancient China. It consists of nearly one hundred scholar cases, each of which records the biographies and academic contributions of a prominent scholar and his disciples.

Nowadays, it is difficult for the younger generation to understand complicated and large-volume Chinese classical writings through word-by-word reading. Meanwhile, modern researchers in humanities also need a knowledge-based visualization tool which can provide an overview of classic materials and help them formulate research questions. Therefor, to make traditional literature more accessible and understandable to both digital natives and humanities researchers, an interactive web-based system, with intuitive visualization functions and assisting tools for background knowledge, is necessary.

In fact, there are several research projects focusing on how to present historical and literary materials. Zhou et al. analyzed Tang poems and visualized the citation

networks among poets(Zhou et al. 2019). Chen et al. built a linked data platform of Chinese Biographical Database and demonstrated the relations between people as a linked graph(Chen et al. 2019). Forlini et al. visualized the Bob Gibson collection of speculative fiction with radial tree graph organized by themes of the fictions(Forlini et al. 2016). Lee et al. analyzed the social networks in Shakespeare's works with diagram(Lee and Lee 2017). Eve created an interactive game to help readers explore the Bible text(Eve 2007).

This paper introduces our work of reconstructing the *Song-Yuan Xue'an* into a knowledge graph upon which several visualization tools are designed for distant reading. To make the classical literature of 300 years ago more comprehensible to history novices, a historical chronology, an ancient bureaucratic hierarchy , and an administrative map of the Song Dynasty are imbedded to provides the temporal, official, and spatial background knowledge respectively for close reading.

2 INTERACTIVE VISUALIZATION TOOLS

First of all, we convert the text of the Song-Yuan Xue'aninto a knowledge graph with two steps. In the first step, the scholars' profiles, spatial-temporal data, and semantic relationships among them are recognized and extracted with regular expressions according to th e structure of the text. Then, all the extracted entities are encoded into RDF triples according to the identified semantic relationships, and then stored into the Neo4j graph database.

We then implement three kinds of visualization tools to demonstrate the development and genealogy of the philosophy schools mentioned in the book. The first one is a streamgraph, representing the varying trend of scholar numbers over the time span of the Song and Yuan Dynasties. The second is a sankey diagram, illustrating the lineage relationships between a Neo-Confucianism school and those derived from it. The third is a pie chart where the center represents a specific scholar and three colored slices represent three group of people related to him: his tutors, disciples and colleagues respectively. All of these visualization tools are interactive, and users can traverse iteratively and explore the underneath knowledge graph.

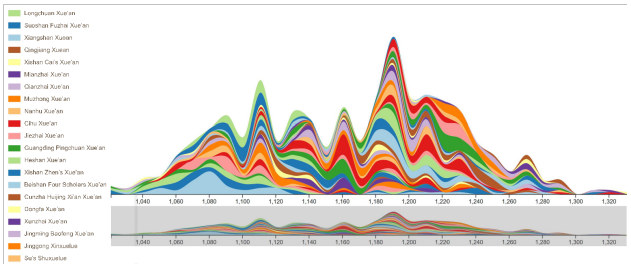


Figure 1: Streamgraph of the number of Scholars in Song and Yuan Dynasties

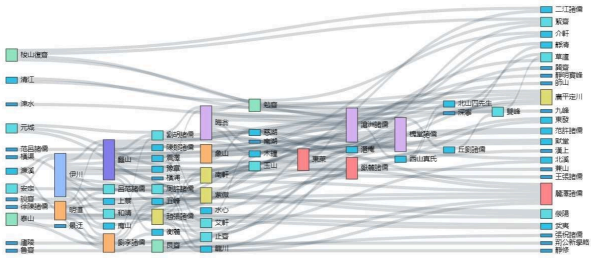


Figure 2: Sankey diagram illustrating lineage relationships among philosophy schools

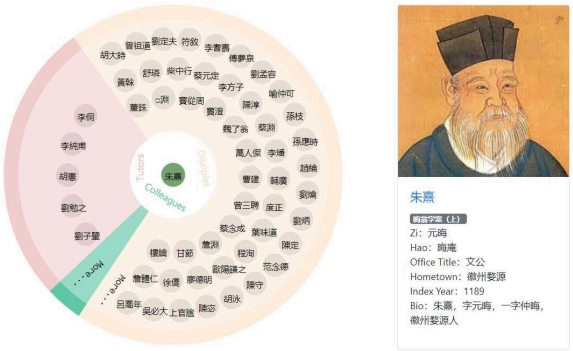


Figure 3: Pie chart demonstrating a scholar's academic circle

3 SYSTEM FUNCTIONALITY

We implement four functions to assist in reading this comprehensive classical work: general search, distant reading, close reading, and semantic search. The general search functions like a search engine. The distant reading function contains three diagrams that provide an overview of *SongYuan Xue'an*. As mentioned above, the streamgraph demonstrates the evolution trend of the NeoConfucian schools over 400 years; and the sankey graph, together with a genealogy tree diagram, shows their origination and influences among each other. The interactive pie chart illustrates the persons related to a particular scholar. The close reading section re-organizes the original text with the semantic relationships extracted from the book, thus related entities of various types are aggregated on the subject that readers are paying attention to. For example, when a person is selected, the places he visited, the book titles he wrote and the official positions he ever took are retrieved from the knowledge graph automatically and displayed alongside his biology. The readers can invoke the embedded historical chronology,

the official hierarchy, and the administrative map of Song Dynasty anytime when he has difficulties in understanding. The semantic search section provides a user-friendly interface to construct complicated semantic queries step by step to search exhaustively in the graph database.

4 CONCLUSION

How to make classic literature more accessible and understandable for both Internet generation and professional researchers is an important question in the pedagogy of the humanities. This project utilizes the knowledge graph and visualization techniques to reconstruct the classic work and make it comprehensible to touch-screen users, demonstrating the value of visualizations in digital humanities fields as a concordance-like tool for classics. The knowledge nodes, including persons, places, time periods, book titles, official titles are extracted from text and represented in RDF triples. With the auxiliary of the embedded hierarchical knowledge schemes, all the knowledge nodes can be aggregated in different granularity interactively to support the function switching from the distant reading to close reading. The coarse-grained knowledge nodes help form a bird-view comprehension of the huge-volume book, and the fine-grained knowledge nodes organizes the text in semantically related unit for close reading and supports complicate semantic search.

For future work, we plan to represent the whole Confucianism history from pre-Qin period to Qing Dynasty in knowledge graph, thus providing on the Internet a vivid interactive picture of Chinese Confucian lineage over two thousand years.

References

Chen, T., Liu, W., Shan, R., and Zhu, Q. (2019). Research on the application of knowledge graph in digital humanities. Journal of Library Science in China, (06):1–19.

Eve, E. (2007). All hope abandon: Biblical text and interactive fiction. Digital Humanities Quarterly, 1(2).

Forlini, S., Hinrichs, U., and Moynihan, B. (2016). The stuff of science fiction: an experiment in literary history. Digital Humanities Quarterly, 10(1).

Lee, J. and Lee, J. (2017). Shakespeare's tragic social network; or why all the world's a stage. Digital Humanities Quarterly, 11(2).

Zhou, L., Hong, L., and Gao, Z. (2019). Construction of knowledge graph of chinese tang poetry and design of intelligent knowledge services. Library And Information Service, 63(2):24–33.

Circulating and Computing Anti-Colonialism: Modeling Revolutionary Information Networks and Radical Ideas in the Long 1960s

Zoe Genevieve LeBlanc
zleblanc@princeton.edu
Princeton University, United States of America

Keywords: information, networks, machine learning, international, anticolonialism
Topics: Comparative (2 or more geographical areas), Global, English, 20th Century, cultural analytics, text mining and analysis, History, Media studies

This paper presents a digital and computational approach to studying the historical of international anti-colonialism in the long 1960s. In particular, I utilize these methods to both craft new datasets and new understandings of how this global movement was shaped by local concerns and linguistic divisions. From Afro-Asianism to the Black Press, I argue that computing anti-colonial discourses can provide new insights into how ideas and information circulated in the emerging Third World.

Linked Archives: a Digital Approach to Visualising Historical Literary Networks

Helen Katherine Bones
h.bones@westernsydney.edu.au
Western Sydney University, Australia

Keywords: digital archives, Australasia, book trade, literary networks
Topics: Global, English, Australia/Oceania, 20th Century, linked (open) data, manuscripts description, representation, and analysis, Book and print history

'Linked Archives' is a digital tool developed to harness the untapped potential of large manuscript collections. Using structured metadata based on Linked Open Data standards, the tool facilitates a range of research techniques relating to documentary evidence and international correspondence networks. Originally designed to describe the business archive of Australian publisher Angus & Robertson, it is particularly suited to studies of transnational literary and publishing operations (such as connections between Australia and New Zealand in the early twentieth century). This paper will explore how 'Linked Archives' can contribute to a critical investigation of literary archives and the stories they tell-or, if possible, the stories that were obscured or not told as the result of these processes.

DH and the Gig Economy: The opportunities and ethical challenges of Sourcery

Sara Sikes
sara.sikes@uconn.edu
University of Connecticut, United States of America

Clarissa Ceglio
University of Connecticut, United States of America

Wes Hamrick
University of Connecticut, United States of America

Tom Scheinfeldt
University of Connecticut, United States of America

Keywords: sharing economy, gig economy, labor, archives
Topics: English, North America, Contemporary, digital research infrastructures development and analysis, mobile applications development and analysis, Humanities computing, Library & information science

Sourcery, a document delivery app, connects researchers around the globe with on-site researchers at some of the world's richest collections of primary sources. But using a sharing economy app to solve problems for researchers elicits now well-rehearsed critiques of the "gig" economy. Recognizing issues of labor equity and responsible regulation, Sourcery will apply the lessons (positive and negative) and technologies of the sharing economy to meet the needs and concerns of both humanities researchers and archives professionals.

Designing accessible platforms: Site-specific digital literature for people with visual impairment

Wilhelm Gustaf Tempelhoff
mrgustaft@gmail.com
North-West University, South Africa

Keywords: Interface design, Accessibility, Site-specific digital literature, Participative inquiry
Topics: Africa, English, Contemporary, electronic literature production and analysis, Interface design, development, and analysis, Communication studies, Design studies

In this paper I discuss the approach and process involved in the design and development of an accessible platform and user interface for people with visual impairment to experience site-specific digital literature. A participative approach, departing from the principle "nothing about us, without us", laid the foundation for the project in collaboration with a school for learners with visual impairment in South Africa. Principles within the fields of universal, accessibility and user centred design, were applied in the development and design of the project. The research and design processes included experimentation with both braille and digital interfaces. In order to facilitate users with various levels of vision I designed and implement a multi-layered user interface consisting of graphical, audio and physical interfaces. With more exploration, the design solutions used in this project could offer new possibilities for accessible spaces, as well as alternative publication solutions for people with visual impairment.

Long Presentations

What is a Genre? A Graph Unified Model of Categories, Texts, and Features

José Calvo Tello
jose.calvo@uni-wuerzburg.de
Göttingen State and University Library

Keywords: genre, model, graph, novels, Spanish
Topics: Comparative (2 or more geographical areas), Europe, English, BCE-4th Century, 19th Century20th Century, cultural analytics, text mining and analysis, Literary studies

Several theoretical models have been proposed for genre, such as the Aristotelian scholastic taxonomy, the family resemblance and the prototype theory. However, these models lack of empirical applications to real examples of genres. This proposal is the culmination of a series of analysis, presenting a theoretical, computational and visual graph-based model that fits several observations. This formalization unifies components of the previous theories, offering visually the intention (internal features) and extension (the best representatives and instances) of each category. Besides, it allows two intuitive interpretations based on the evaluation: the centrality as classification results, and the distance as similarity through shared features. The model is applied to three data-sets of different periods and languages: modern Spanish novels, classic French plays and the books of the Bible.

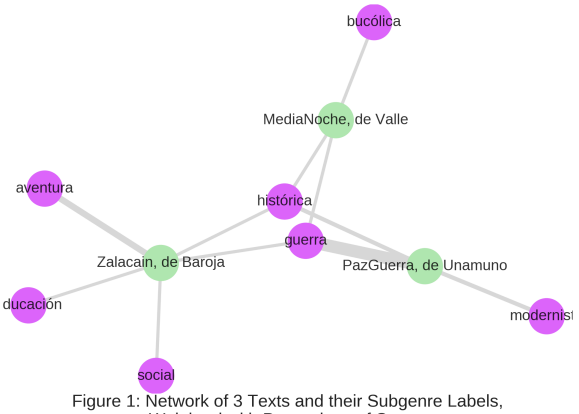


Figure 1: Network of 3 Texts and their Subgenre Labels, Weighted with Proportions of Source

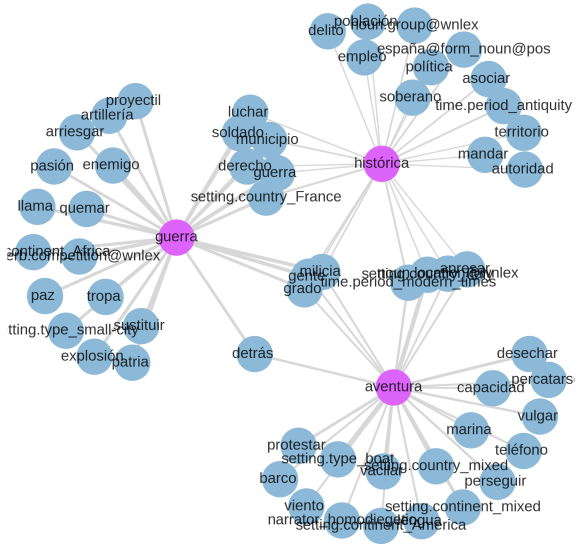


Figure 2: Network of 3 Subgenres and their Distinctive Features Weighted as Z-Scores

Figure 3: Network of Texts, Features and 4 Subgenres (adventure, historical, war and gregueríes)

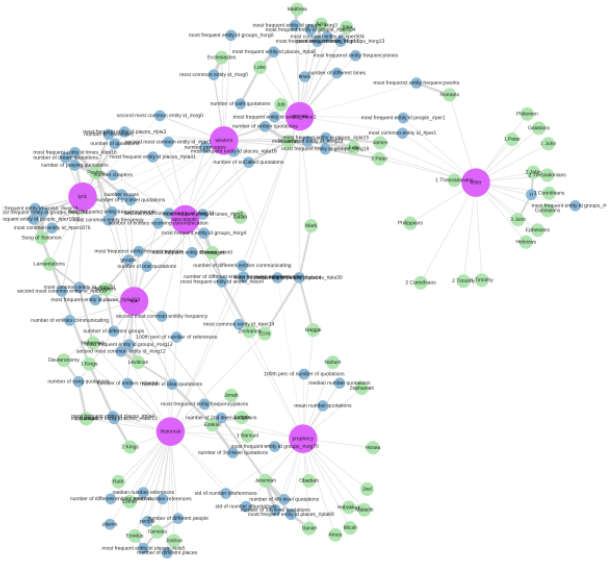


Figure 6: Network of Texts, Features and Genres of the Bibel

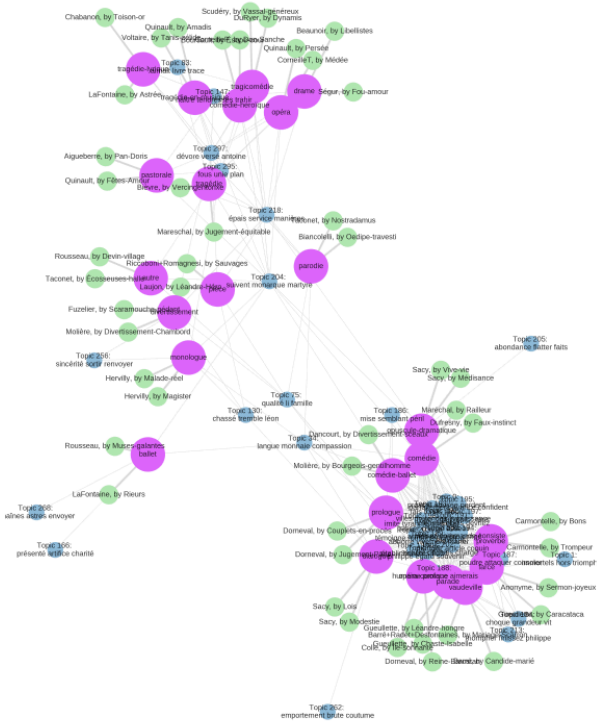


Figure 5: Network of Texts, Topics and Subgenres of French Plays

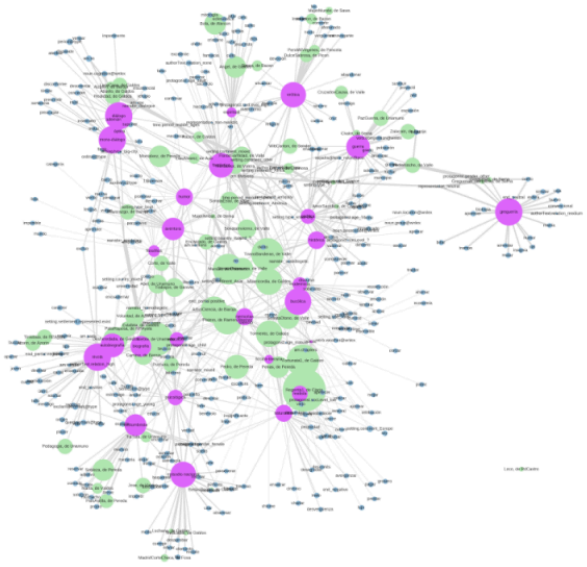


Figure 4: Network of all Subgenres, Texts, and Features of Spanish Novels

Restoration and Repurposing of DH Legacy Projects

Alessio Antonini
alessio.antonini@open.ac.uk
The Open University, United Kingdom

Francesca Benatti
francesca.benatti@open.ac.uk
The Open University, United Kingdom

Edmund King
The Open University, United Kingdom

Keywords: Information Systems, Legacy Systems
Topics: Global, English, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), sustainable procedures, systems, and methods, Informatics

The institutional funding system of Digital Humanities (DH) is usually devoted to the creation of new projects, creating a recurring problem of unsupported legacy projects whose material cost of upkeep depends on the voluntary contributions of institutions and individuals. The lack of resources to invest in "remedial" actions [10] pushes DH projects towards outdatedness. Additional funding success delays this process by introducing extra resources but, simultaneously, it fast-forwards obsolescence by advancing the field.

Indeed, the impact of a DH project can be considered as the ability to establish as common knowledge what was once innovative and cutting-edge by fulfilling its research questions. In this scenario, managing successful DH projects requires addressing competing issues related to the preservation of their integrity [1] (i.e. consistency of data, questions and vision) and of their role and purpose (i.e. their use in the field).

The management of legacy systems has been widely studied from a technical perspective [2,3,4,5,6,7,8], e.g. cost/value [9], approach to integration [1,9], change of use [1] and archiving [3,4]. Rather than presenting technical solutions, this contribution focuses on the rationale for defining an approach through human, financial and political perspectives [2].

The issue of legacy is not one of data formats but principally a cultural one that we analyse from two distinct approaches:

The "restoration" approach, implementing remedial actions [10] that "update" the project to new contexts to preserve its function and role (e.g. extending its data structure to address new questions)

The repurposing approach [1], implementing actions that rethink the value of the project by finding it new purposes, functions and roles in new contexts (e.g. defining new questions to be addressed with its existing data).

At stake in the two approaches are the integrity and identity of the project. A project's integrity is the logical and historical connection between its origin, output and outcomes. A project's identity is the meaning or role it has within the community of people involved. In this scenario, we argue that addressing a project's legacy should take into account:

The project's vision, research questions and target "knowledge gap"; the project creators' motivations and aims; the funding bodies' goals and priorities

The project's practices, orchestration of people, organisations and tools, operational limits and constraints

The project's knowledge, research data and outputs, correlated research activities, answers produced and outcomes (e.g. new projects, scholarly research, education, impact on the field).

The contribution then discusses a real case, the UK Reading Experience Database (RED), as emblematic of the challenges of managing legacy DH projects. RED has had a long history, repeated funding successes and significant visibility in Book History scholarly literature. It was devised by Simon Eliot in 1993, first implemented in 1996, published on the web in 2007 and finally closed to new submissions in 2018 [11]. RED's vision was to advance research in the history of reading by establishing a new methodology based on empirical evidence [12]. RED's practice established strong synergies between researchers, students and volunteers for the distributed acquisition and curation of evidence of reading. The RED contribution form's structured approach to knowledge encouraged data inputters to pay close attention to the contexts and agents involved in the reading experience: who was reading, what was read, and when and where the act of reading took place [13].

RED data has been successfully converted from a legacy custom relational database to linked open data. Still, RED is a legacy project because of what its data expresses about reading experiences: a now-outdated vision established more than twenty years ago, which has now become embedded in the DH community through successful activities, follow-on projects, publications, research and training initiatives [14,15,16,17]. RED is both a vast database and the centre a wide network of collaborations; therefore addressing its legacy is not a trivial decision.

1. The Repurposing of RED: Repurposing RED's vision means, for instance, rethinking its role from research infrastructure to an educational resource. Consequently, RED's practices could be reframed as a playground for DH students, providing an environment for training and annotation evaluation. The knowledge produced and encompassed in RED could document the history of DH methods or become a training set for machine-learning algorithms.

2. The Restoration of RED. Restoring RED's vision means, for instance, incorporating in RED new approaches currently required by funding bodies (e.g. collaboration with data science) and current research priorities within Book History, changed considerably since 1993. Consequently, the study of sources could be combined with machine-learning and natural-language processing tools not included in the original structure of RED. Finally, new research questions such as the effects of reading [18] and the multi-modality of reading on new media [19] could be addressed.

With a repurposing approach, the integrity of the dataset could be preserved by relinquishing RED's role

as a research project. With a restoration approach, RED's role as a research project could be preserved through the entire re-curation of its data, the complete re-development of the tool ecosystem to include automatic steps and the entire reassessment of its value as a research resource in light of current DH and Book History research agendas. Unsurprisingly, to keep a project's role we must face the cost of adapting to the new context, while to keep its form, we must search for a new purpose.

On a more general level, there is a question about how to preserve the "human legacy" of RED, e.g. the network of collaborations, student volunteers and contributors engaged. A DH project is a Cultural Artefact, and therefore its historical context can guide the re-tuning of its role as the context changes or the search for new purposes compatible with the values and vision of the social system of the project.

As a final remark, we hope these questions can elicit a broader discussion about new and future DH projects and how we could design for their legacy, e.g. the new Reading Europe Advanced Data Investigation Tool.

How can science and knowledge be created for all and by all without #linguisticjustice?: Findings from a two-year study on the intersections between citizen science, social media, crowdsourcing, and Translation Studies.

Renée Desjardins
rdesjardins@ustboniface.ca
Université de Saint-Boniface, Canada

Keywords: Translation Studies, linguistic justice, knowledge dissemination, Citizen Science
Topics: Global, English, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), social media analysis and methods, Humanities computing, Translation studies

This paper presents the findings of a 2-year study examining the role of translation and language diversity in online citizen science initiatives. In the last 10 years, the proliferation of mobile technologies, the popularity of participatory culture and social media, as well as the uptick in crowdsourced models for conducting large-scale tasks has impacted the academic landscape. Specifically, this project considers two platforms, Zooniverse and the Canadian Citizen Science Portal. The overarching framework is an amalgamation of methods and theories, situating the project in the transdisciplinary Digital Humanities. Because the project's mandate is, in part, to informmore equitable exchange/dissemination of citizen science capitals in online and digital spaces, the underpinning philosophical worldview is one that is transformative. Within the purview of Translation Studies, this study falls under the umbrella of descriptive product-oriented research and context-oriented research. Social network analysis (data visualization) and social media analysis (qualitative) further supplements this framework.

Algorithms of Resistance: Using OCR and AI for Social Justice

Nathan Kelber
NKELBER@GMAIL.COM
JSTOR Labs, United States of America

Keywords: jim crow, algorithmic bias, ocr, tdm
Topics: English, North America, 19th Century, 20th Century, optical character recognition and handwriting recognition, text mining and analysis, African and African American Studies, Law and legal studies

This long presentation will share key findings from the final report of *On the Books: Jim Crow and Algorithms of Resistance*, a Mellon-funded Collections as Data project using machine learning to systematically discover racism within North Carolina laws. The project will make North Carolina legal history accessible to researchers by creating a corpus of over one hundred years of North Carolina public, private, and local session laws and resolutions from the end of civil war through the civil rights movement (1865-1968).

This project represents the intersection of many subject areas including critical race theory, American history, text and data mining, machine learning, and public humanities. It will be of interest to subject specialists, data science researchers, educators, librarians, and other information professionals.

The place of models and modelling in Digital Humanities: Intersections with a Research Software Engineering perspective

Arianna Ciula
arianna.ciula@kcl.ac.uk
King's College London, United Kingdom

Paul Caton
King's College London, United Kingdom

Ginestra Ferraro
King's College London, United Kingdom

Brian Maher
King's College London, United Kingdom

Geoffroy Noël
King's College London, United Kingdom

Miguel Vieira
King's College London, United Kingdom

Keywords: Research Software Engineering, models, critical modelling
Topics: Comparative (2 or more geographical areas), English, Contemporary, data modeling, project design, organization, management, Humanities computing

This paper[1] aims to bridge Digital Humanities (DH) and Research Software Engineering (RSE) communities. It argues that the production of models is the core contribution of RSE to the epistemology of DH. We adopt an inclusive definition of models and modelling (see Ciula et al. 2018) which spans the whole range from 'deformative' to empirical modelling (see Smithies 2017: 168), including formal or predictive modelling (Joslyn and Turchin 1993), and the technical solutions produced in the process as well as the know-how, languages and documentation which accompany this production. From this wide perspective, models are also artefacts which can be studied across the history of science and of the humanities tradition (see Bod 2018) and in comparison with other modelling practices in science. RSE practice is grounded on a strong conscience of the experimental apparatus and the iterative critique of models built (and often deflated) for a purpose. The challenge is to recognise the idiosyncrasy and situatedness of modelling practices and artefacts while devising methods to expose the scalability of the underlying workflows and modelling processes.

We reflect on the epistemology of DH from the practical perspective of our RSE lab - King's Digital Lab (KDL)[2] - and the research processes embedded in our Software Development Lifecycle (fig.1). The human element is at the core of the technical ecosystem we research and operate in. We acknowledge that KDL models and modelling are co-constitutive of human expertise, technical systems and operational methods, all aspiring to an environment conducive of open knowledge. This is not only in terms of development and management approaches via adoption of open standards, open source and exposure of open data but also, more fundamentally, in sharing (achievements and struggles around) our processes and in promoting open models. We will use project-specific examples, including data modelling and knowledge representation practices, to demonstrate how some of our research into model-making processes challenge the perception of the technical work of RSE within DH as a stale, mechanistic and uncritical procedural activity.

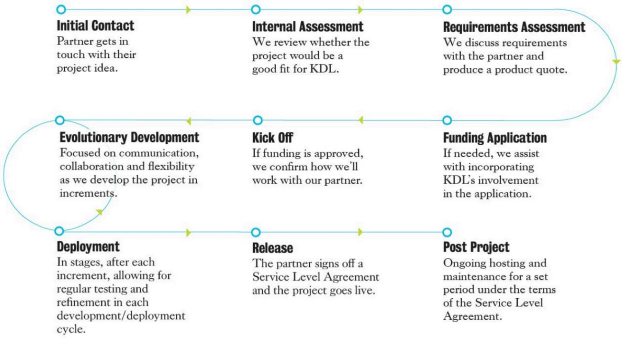


Figure 1 KDL Software Development Lifecycle (SDLC) (King's Digital Lab 2019) mapped to Agile DSDM project phases (Agile Business Consortium 2014).

KDL operates within a rather unique context. It claims its origins in the pioneering work of colleagues at King's College London working in applied computing in the Humanities already in the 1970s (see Short et al. 2012). However, the crossings between RSE and DH communities at King's and internationally are only recently being highlighted and explored (e.g. Gold 2009; Smithies 2019; DHTech Group 2019). We argue that the study and building of models is one of the dimensions via which these crossings emerge more vividly with substantial epistemological implications and innovative ramifications for the field of DH as a whole.

The core practice of research in DH is modelling (e.g. cf. McCarty 2005: 20-72; Buzzetti 2002; Beynon and McCarty 2006; Flanders and Jannidis 2015, 2018), which implies the translation of complex systems of knowledge or conceptual frameworks into computationally processable models or operational frameworks.

Gooding (2003) claims that in experimental settings computational approaches are analogues to other processes of abstraction, measurement and contextual interpretation, whereby reduction of complexity is followed by expansion in the guise of a double funnel-shaped process (Gooding 2003: fig 13.4). We can trace these processes of reduction and expansion also in the RSE context, where, for example, operationalisation makes models formalised into snippets of code or software components. Other languages than translation into code play a role in the process, however.[3]

The paper will reflect on models as artefacts of different kind expressed via a variety of languages, including but not limited to computational models, produced during several phases of the SDLC (see Ciula and Smithies, forthcoming) such as:

- negotiations around the meaning of the project units of analysis documented in diagrams and definitions which shape requirements and an agreed project language;
- paper or whiteboard sketches used to draft the solution architecture for a project in its feasibility assessment;
- wireframes and static mockups of user journeys;
- data models implementing the logical structure of a database;
- statistical models implemented with ad hoc algorithms and code or relying on tested formulas and existing libraries.

While specific instantiations of models (for example in relational databases) can have a rather short life, in the RSE context and projects where they were designed and developed, they often represented innovative solutions which have a longstanding effect. Indeed, while models are temporary pragmatic solutions to address specific project challenges yet, at the scale KDL operates, they are the backbone of the team's tacit knowledge as well as the building blocks towards more generalisable and re-usable approaches. They mediate and bridge the layers of the Lab's socio-technical system: team expertise, data and technical systems (fig. 2).

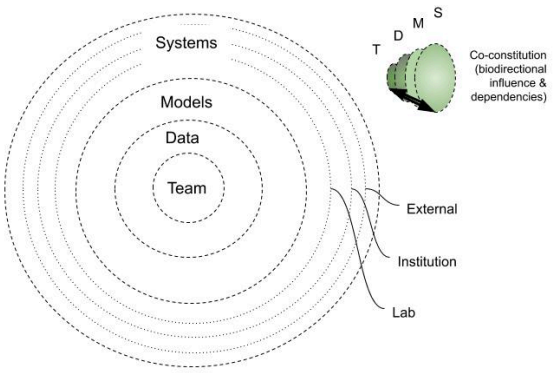


Figure 2 Multilayered socio-technical system of the Lab where concentric circles denote co-constitution of team expertise, data and technical systems (Ciula and Smithies, forthcoming: fig. 5).

While we can refine existing approaches to sustain and expose modelling efforts in SDLC cycles which rely on RSE best practices such as attention to documentation and re-use,4 there is also room for more innovative approaches, including a design first culture, workflow integration across RSE roles, the assessment and potential adoption of a set of modelling notation languages for the exposure mechanisms for our models.

In alignment with a "critical modelling" approach (see Bode 2020), but also with a material culture and media literacy perspective, in this paper we aim to reflect upon models by looking at the wide epistemological implications of their production and use, at the responsibilities of modellers,5 at how models come to be and what effect they have in the resources they contribute to instantiate and hence in interpretative processes of expansion as well as reduction.

Notes

[1] A preliminary version of this paper was presented at the symposium Computational Text Analysis and Historical Change, held at Humlab, Umeå University (Sweden), 4-6 September 2019.

[2] The authors currently cover different roles at King's Digital Lab (KDL), a Research Software Engineering (RSE) team hosted by the Faculty of Arts & Humanities at King's College London, which provides software development and infrastructure to

departments in the Faculty of Arts & Humanities while also collaborating with Social Science & Public Policy as well as a range of external partners in the higher education and cultural heritage sectors.

[3] Data modelling, for example, is informed if not driven by communication and collaborative reasoning around more or less standardised graphical representations and notations in phases of reverse engineering as well as design methods. Note that KDL intend and use design methods in a wide sense ranging from techniques of requirements elicitation in pre-project analysis to data modelling and wireframing in evolutionary development (see Bennet et al. 2005). Equally, the re-integration or expansion of modelling efforts into interpretative frameworks usually rely on verbal and visual language to document code, or to explain the results of an experiment (Ciula and Marras 2019: 39).

[4] With this respect see, for example, the KDL checklist for assessment of digital outputs within the UK Research Framework Exercise (Ciula 2019).

[5] Models contribute to define and redefine objects of study which come charged with layers of scholarship and analysis, with previous selections, bias and political as well as ethical responsibilities. As the creators of new memory regimes and intermediaries to the past engaged in modelling efforts which interact and affect the materiality of our objects of study (Ciula 2017a), we bear responsibilities (Ciula 2017b). In line with ongoing discussions around the representativeness and constraints of the digital archive (e.g. Dahlström 2010; Hitchcock 2013; Prescott 2015), Bode (2018, 2020) presented some lucid analysis around modellers' responsibilities in digital literary studies by exposing the gaps that propagate from produced literary works we know of to material preserved in the analogue archive, to selections of works that make it into digital archives to further reductions in the creation of a corpus of analysis and, last but not least, in the application of statistical modelling techniques which dictate additional powerful yet limiting constraints if not contextualised critically within an interlocked chain of bias.

References

Agile Business Consortium. The DSDM Agile Project Framework Handbook. DSDM Consortium, 2014. https://www.agilebusiness.org/resources/dsdm-handbooks.

Bennett, Simon, Steve McRobb, and Ray Farmer. Object-Oriented Systems Analysis and Design Using UML. McGraw Hill Higher Education, 2005.

Beynon, Meurig, Steve Russ, and Willard McCarty. 'Human Computing-Modelling with Meaning'. Literary and Linguistic Computing 21, no. 2 (6 January 2006): 141-57. https://doi.org/10.1093/llc/fql015.

Bod, Rens. 'Modelling in the Humanities: Linking Patterns to Principles'. Historical Social Research, Supplement, no. 31 (2018): 78-95. https://doi.org/10.12759/hsr.suppl.31.2018.78-95.

Bode, Katherine. A World of Fiction: Digital Collections and the Future of Literary History. University of Michigan Press, 2018.

---. 'Why You Can't Model Away Bias'. Modern Language Quarterly 81, no. 1 (1 March 2020): 95-124. https://doi.org/10.1215/00267929-7933102. preprint: https://katherinebode.files.wordpress.com/2019/08/mlq2019_preprintbode_why.pdf.

Buzzetti, Dino. 'Digital Representation and the Text Model'. New Literary History 33, no. 1 (2002): 61-88. https://doi.org/10.1353/nlh.2002.0003.

Ciula, Arianna. 'Digital Humanities and Practical Memory: Modelling Textuality'. Estudos Em Comunicação 25 (21 December 2017): 7-17.

---. 'Modelling Textuality: A Material Culture Framework'. In Advances in Digital Scholarly Editing. Papers Presented at the DiXiT Conferences in The Hague, Cologne, and Antwerp, edited by Edited by Peter Boot, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Andreas Mertgens, Anna-Maria Sichani, and Elena Spadini & Dirk van Hulle, 91-97. Leiden: Sidestone Press, 2017. https://www.sidestone.com/books/advances-in-digital-scholarly-editing.

---. 'KDL Checklist for Digital Outputs Assessment (Version 2.0)'. Zenodo, 2019. http://doi.org/10.5281/zenodo.3361580.

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle, eds. 'Models and Modelling between Digital & Humanities: A Multidisciplinary Perspective'. Historical Social Research Supplement, no. 31 (2018). https://doi.org/10.12759/hsr.43.2018.4.343-361.

Ciula, Arianna, and Cristina Marras. 'Exploring a Semiotic Conceptualisation of Modelling in Digital Humanities Practices'. In Meanings & Co: The Interdisciplinarity of Communication, Semiotics and Multimodality, edited by Alin Olteanu, Andrew Stables, and Dimitru Bortun, 6:33-52. Numanities - Arts and Humanities in Progress. Springer International Publishing, 2019. https://www.springer.com/us/book/9783319919850.

Ciula, Arianna, and James Smithies. 'Sustainability and Modelling at King's Digital Lab: Between Tradition and Innovation'. In On Making in the Digital Humanities: Essays on the Scholarship of Digital Humanities Development in Honour of John Bradley, edited by Julianne Nyhan, Geoffroy Rockwell and Stefan Sinclair. London: University College London Press, Forthcoming.

Dahlström, Mats. 'Critical Editing and Critical Digitisation'. In Text Comparison and Digital Creativity:

The Production of Presence and Meaning in Digital Text Scholarship, edited by Wido van Peursen, Ernst D. Thoutenhoofd, and Adriaan van der Weel, 79-97. Brill Academic Pub, 2010.

DHTech Group. 'DH Research Software Engineers - For We Are Many'. DHTech, 2019. https://dh-tech.github.io/dhrse-whitepaper/.

Flanders, Julia, and Fotis Jannidis. 'Knowledge Organization and Data Modeling in the Humanities'. White paper, 2015. http://www.wwp.northeastern.edu/outreach/conference/kodm2012/flanders_jannidis_datamodeling.pdf.

---, eds. The Shape of Data in Digital Humanities: Modeling Texts and Text-Based Resources. Digital Research in the Arts and Humanities. London: Routledge, Taylor & Francis Group, 2018.

Gold, Nicolas. 'Service-Oriented Software in the Humanities: A Software Engineering Perspective'. Digital Humanities Quarterly, no. 3.4 (2009). http://digitalhumanities.org:8081/dhq/vol/3/4/000072/000072.html.

Gooding, David. 'Varying the Cognitive Span: Experimentation, Visualization, And Computation'. In The Philosophy of Scientific Experimentation, edited by Hans Radder, 255-84. Pittsburgh, Pa.: University of Pittsburgh Press, 2003.

Hitchcock, Tim. 'Confronting the Digital'. Cultural and Social History 10, no. 1 (1 March 2013): 9-23. https://doi.org/10.2752/147800413X13515292098070.

Joslyn, Cliff A., and Valentin Turchin. 'Model'. In Principia Cybernetica Web, edited by Francis Heylighen, Cliff A. Joslyn, and Valentin Turchin. Brussels, 1993. http://pespmc1.vub.ac.be/MODEL.html.

King's Digital Lab. 'Frequently Asked Questions for Project Partners | King's Digital Lab'. King's Digital Lab, 2019. https://www.kdl.kcl.ac.uk/how-we-work/faq-partners/.

---. 'What Is KDL | King's Digital Lab'. King's Digital Lab, 2018. https://www.kdl.kcl.ac.uk/how-we-work/what-is-kdl/.

McCarty, Willard. Humanities Computing. Basingstoke [England]; New York: Palgrave Macmillan, 2005.

Prescott, Andrew. 'I'd Rather Be a Librarian: A Response to Tim Hitchcock, "Confronting the Digital"'. Cultural and Social History 11, no. 3 (1 September 2014): 335-41. https://doi.org/10.2752/147800414X13983595303192.

Short, Harold, Julianne Nyhan, Anne Welsh, and Jessica Salmon. '"Collaboration Must Be Fundamental or It's Not Going to Work": An Oral History Conversation between Harold Short and Julianne Nyhan'. Digital Humanities Quarterly 6, no. 3 (2012). http://www.digitalhumanities.org/dhq/vol/6/3/000133/000133.html.

Smithies, James. 'The Continuum Approach to Career Development: Research Software Careers in King's Digital Lab'. King's Digital Lab - Thoughts and Reflections from the Lab (blog), 7 February 2019. https://www.kdl.kcl.ac.uk/blog/rse-career-development/.

---. The Digital Humanities and the Digital Modern. Basingstoke: Palgrave Macmillan, 2017.

The Open Knowledge Program: Creating Space for Digital, Public Scholarship

Caroline Winter
winterc@uvic.ca
Electronic Textual Cultures Lab, University of Victoria

Randa El Khatib
khatib@uvic.ca
Electronic Textual Cultures Lab, University of Victoria

Alyssa Arbuckle
Electronic Textual Cultures Lab, University of Victoria

Ray Siemens
Electronic Textual Cultures Lab, University of Victoria

Keywords: open scholarship, open knowledge, social knowledge creation, digital humanities, public humanities
Topics: English, North America, Contemporary, digital research infrastructures development and analysis, public humanities collaborations and methods, Humanities computing

DH2020 Proposal
Caroline Winter
Randa El Khatib
Alyssa Arbuckle
Ray Siemens
The Open Knowledge Program: Creating Space for Digital, Public Scholarship

In *Generous Thinking: A Radical Approach to Saving the University* (2019), Kathleen Fitzpatrick argues that

academics must reconsider themselves within the "larger 'us' that we together form," rather than holding themselves apart from the wider community (8). How to enact more open, public work is not always obvious, however. Researchers often face barriers to engaging in open and public-facing scholarship, including lack of training, infrastructure, and technical and community support. The Open Knowledge Program at the University of Victoria's Electronic Textual Cultures Lab (ETCL) facilitates intersections between the scholarly and public communities by supporting university and community researchers in creating open knowledge: "what open data becomes when it's useful, usable and used" (Open Knowledge Foundation n.d.). In this paper, we discuss the trajectory of the program so far, share examples of participants' contributions, and invite feedback and discussion about adapting the program for other contexts and its next steps.

The Program

The Open Knowledge Program is based in the ETCL, a collaborative humanities research lab with a focus on open scholarship and a mandate that includes research, training, and service.

The Program comprises three initiatives: the Open Knowledge Practicum (OKP), the Open Knowledge Practicum at the Digital Humanities Summer Institute (OKP@DHSI), and the Open Knowledge Residency (OKR). The OKP is a term-long program that welcomes researchers from UVic and the wider community into the lab to work on an open knowledge project, including contributing to Wikipedia. The OKP@DHSI supports the ETCL's global community by inviting DHSI students and instructors into the lab for a condensed, three-day version of the OKP. The OKR offers graduate students in any discipline at UVic an intensive, week-long residency in the lab to conduct thesis or doctoral research and share findings in an open venue.

The Intervention

Just as digital scholarship has moved from the periphery of the Humanities towards its centre, open scholarship is increasingly recognized as the new scholarly mode. This transition from closed to open is driven partly by necessity - the costs of the current subscription model of scholarly communication being unsustainable for many research libraries - and partly by researchers driven to engage with the broader community and universities striving to fulfill their public missions (CARL-ABRC 2010; O'Gara 2019; Suber 2019). It is also driven by opportunity, since digital technologies have made it possible to share scholarly work widely with academic and public readers, greatly extending its "reach" (Maxwell 2015, 2).

Digital tools allow researchers to engage with new types of research materials, new tools and methodologies, and new modes of communication, but require skills that are not part of traditional Humanities curricula, including collaboration. These digital skills are increasingly recognized as essential for scholarly work and beyond, particularly for emerging scholars (Brier 2012; El Khatib, Arbuckle, and Siemens 2019; IPLAI 2013; Jakacki and Faull 2016; Lewis et al. 2015; MLA 2014; NEH 2016; Reid 2012). The Open Knowledge Practicum draws on the pedagogical model of the practicum, in which skills and knowledge are applied in practice. Practicums are common in Education and professional fields including nursing and clinical psychology, but much less so in the Humanities and Social Sciences. Although practicums are usually part of a larger curriculum, the OKP employs this model as a standalone program, in which researchers plan, develop, and create their own research projects. In doing so, participants put their subject matter expertise and digital skills into action by creating open knowledge resources while gaining experience working in a DH lab as part of a research team. Participants develop their own projects in consultation with the lab team and our colleagues in the Library and across campus as needed. The lab team also provides just-in-time collaborative learning as needed in the use of specific software applications, digital tools, and project management. Honorariums for participants comprise registration in a DHSI course, providing further opportunity for learning, putting learning into practice, and building community.

In its focus on open scholarship as praxis, the Open Knowledge Program complements the "apprenticeship model that dominates graduate training and socialization in the humanities and elsewhere across the university," which focuses tightly on skills necessary to institutional frameworks (Bartha and Burgett 2015, 33-4). These frameworks, however, leave little space for digital and public scholarship: review, promotion, and tenure guidelines, for example, tend to discourage open scholarship in favour of more traditional forms of scholarly communication (Alperin et al. 2018). In their study of best practices for supporting digital scholarship, Lewis, Spiro, Wang, and Cawthorne note that physical space in which to work, collaborate, and learn is key (2015, 2). Lewis et al. also find that "collaborative competencies" and "learning mindsets," comprising "creativity, curiosity, and an enthusiasm for learning" are as important for digital scholarly work as technical skills (2015, 2). Building regular lab hours into the Program's structure provides participants with a shared physical workspace as well as a highly collaborative community of practice that facilitates social knowledge creation (Burke 2000). Participants in the Program determine the subject and scope of their project, pursuing interests that intersect with

their academic work or study, or not. In this way, the Program creates intellectual space and scheduled time for curiosity, exploration, and creativity, space that is often difficult to find within a university's institutional structure (Bartha and Burgett 2015). Many projects focus on social justice issues such as wealth inequality, Indigenous knowledge systems, international LGBTQ rights, and histories of oppressed groups (ETCL 2020; El Khatib et al. 2019).

The Open Knowledge Program has evolved over the past few years to support standalone projects as well as multi-term endeavours and components of large, grant-funded research initiatives with faculty partners and their student research assistants. We anticipate that describing the structure and goals of the Open Knowledge Program and highlighting the work of its participants will open a discussion of the Program's future directions and provide a model for other digital scholarship centres that are invested in public scholarship and open knowledge production.

References

Alperin, Juan Pablo, Carol Muñoz Nieves, Lesley Schimanski, Gustavo E. Fischman, Meredith T. Niles, and Erin C. McKiernan. 2018. "How Significant are the Public Dimensions of Faculty Work in Review, Promotion, and Tenure Documents?" Humanities Commons [preprint]. http://dx.doi.org/10.17613/M6W950N35.

Bartha, Miriam, and Bruce Burgett. 2015. "Why Public Scholarship Matters for Graduate Education." Pedagogy 15, no. 1: 31-43. https://doi.org/10.1215/15314200-2799148.

Brier, Stephen. 2012. "Where's the Pedagogy? The Role of Teaching and Learning in the Digital Humanities." In Debates in Digital Humanities, edited by Matthew K. Gold, 390-401. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/9781452963754.

Burke, Peter. 2000. A Social History of Knowledge: From Gutenberg to Diderot. Cambridge: Polity Press.

CARL-ABRC (Canadian Association of Research Libraries-Association des bibliothèques de recherche du Canada). 2010. CARL Open Access Backgrounder. http://www.carl-abrc.ca/doc/CARL_OA_Backgrounder_EN.pdf.

El Khatib, Randa, Alyssa Arbuckle, and Ray Siemens. 2019. "Foundations for On-Campus Open Social Scholarship Activities." KULA: Knowledge Creation, Dissemination, and Preservation Studies 3, no. 1. http://doi.org/10.5334/kula.14.

El Khatib, Randa, Lindsey Seatter, Tracy El Hajj, Conrad Leibel, Alyssa Arbuckle, Ray Siemens, Caroline Winter, and the ETCL and INKE Research Groups. 2019. "Open Social Scholarship Annotated Bibliography." KULA: Knowledge Creation, Dissemination, and Preservation Studies 3, no. 1. http://doi.org/10.5334/kula.58.

ETCL (Electronic Textual Cultures Lab). 2020. "Open Knowledge Practicum." https://etcl.uvic.ca/open-knowledge-practicum-okp/.

Fitzpatrick, Kathleen. 2019. Generous Thinking: A Radical Approach to Saving the University. Baltimore: Johns Hopkins University Press.

IPLAI (Institute for the Public Life of Arts and Ideas). 2013. "White Paper on the Future of the PhD in the Humanities." 2013. Institute for the Public Life of Arts and Ideas, McGill University. http://iplai.ca/wp-content/uploads/2015/04/white_paper_on_the_future_of_the_phd_in_the_humanities_dec_201 3

Jakacki, Diane, and Katherine Faull. 2016. "Doing DH in the Classroom: Transforming the Humanities Curriculum through Digital Engagement." In Doing Digital Humanities, edited by Constance Crompton, Richard Lane, and Ray Siemens, 358-72. London: Routledge.

Lewis, Vivian, Lisa Spiro, Xuemao Wang, and Jon E. Cawthorne. 2015. Building Expertise to Support Digital Scholarship: A Global Perspective. Council on Library and Information Resources. https://www.clir.org/pubs/reports/pub168/.

Maxwell, John W. 2015. "Beyond Open Access to Open Publication and Open Scholarship." Scholarly and Research Communication 6, no. 3. https://doi.org/10.22230/src.2015v6n3a202.

MLA (Modern Language Association). 2014. Report of the Task Force on Doctoral Study in Modern Language and Literature. https://www.mla.org/content/download/25437/1164354/taskforcedocstudy2014.pdf.

NEH (National Endowment for the Humanities). 2016. "NEH Announces $1.7 Million for 'Next Generation PhD.'" August 9, 2016. https://www.neh.gov/news/press-release/2016-08-10.

O'Gara, Genya, and Anne C. Osterman. 2019. "Negotiating on Our Terms: Harnessing the Collective Power of the Consortium to Transform the Journal Subscription Model." Collection Management 44, no. 2-4: 176-194. https://doi.org/10.1080/01462679.2018.1564716.

Open Knowledge Foundation. n.d. "What is Open?" https://okfn.org/opendata/. Accessed October 9, 2019.

Reid, Alexander. 2012. "Graduate Education and the Ethics of the Digital Humanities." In Debates in the Digital Humanities, edited by Matthew K. Gold, 350-67. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/9781452963754.

Suber, Peter. 2019. "2. Motivation." In Open Access, by Suber, https://openaccesseks.mitpress.mit.edu/pub/ktf344br.

What's in a Face? Gender representation of faces in *Time*, 1940s-1990s

Ana Jofre
jofrea@sunypoly.edu
SUNY Polytechnic, United States of America

Josh Cole
acole3@gmail.com
Queen's University, Canada

Michael Reale
SUNY Polytechnic, United States of America

Vincent Berardi
Chapman University, United States of America

Keywords: Time Magazine, Gender representation, Image analysis
Topics: English, North America, 20th Century, cultural analytics, image processing and analysis, Computer science, History

Introduction

Beginning with its inception in 1923, Time magazine, perhaps more than any comparable publication, has both reflected and influenced American popular attitudes to domestic and global politics. These include the changing ideas about women since the mid-twentieth century, which is the subject of this paper.

We used supervised machine learning to extract 327,322 visual images of faces from an archive of Time magazine, which contains 3389 issues ranging from 1923 to 2014, and computationally classified the faces as male or female. We then closely read selected Time articles to make sense of this quantitative data against the background of postwar feminism writ-large, and the history of the magazine itself. Our focus is on the period between the 1940s and the 1990s, which witnessed significant changes in attitudes toward women, and where our data of the proportion of female faces exhibits significant fluctuation.

We found four clear phases in the visual representation of women in Time from the 1940s to the 1990s: a peak in the mid-to-late 1940s, a dip from the mid-1950s to early 1960s, another peak in the 1970s, and another dip in the 1980s. The number of female faces depicted in Time then rises steadily since the early-1990s. We interpret these variations through an interdisciplinary framework. Through our combined quantitative and qualitative approach, we found that the percentage of female faces found in Time between 1940 and 1990 correlates with attitudes towards women in both the larger historical context as well as within the textual content of the magazine.

Methods and Results

We first collected data through human labor using Amazon Mechanical Turk (AMT) to identify and tag faces from the archive. This data was used to train a RetinaNet detector[1] to automatically identify and extract faces from the remainder of the archive. Using an accuracy threshold of 90 percent yielded 327,322 faces. A pre-trained face descriptor convolutional neural network VGGFace[2] was then fine-tuned and used to classify each face as either male or female.

Similar patterns emerge from both the AMT data and the automated data, featuring an increase in the proportion of female faces from the 1920s to 1945, a post-Second World War dip, a rebound beginning in the mid-1960s, followed by a decrease in the 1980's, and a final rebound beginning in the early-1990s. Since similar trend lines were found using the AMT and the automatic extraction data, our analysis focuses on the latter, more comprehensive data set.

The proportion of women in each issue is shown in Figure 1. While there is a significant amount of variance per issue (compared to when the data is aggregated per year), a clear trendline emerges when the data is Lowess smoothed. Charting the proportion of women in each issue was useful for identifying outliers for our close reading analysis.

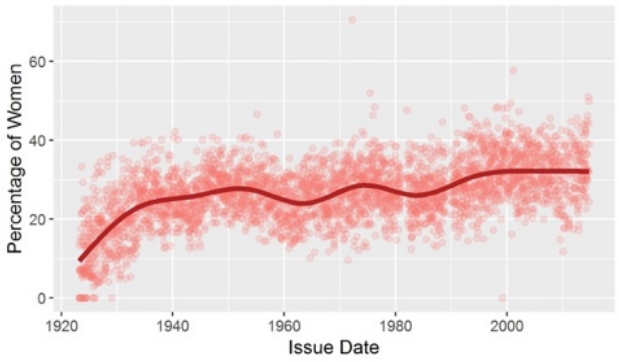


Figure 1: The percentage of women's faces in each issue. The solid line is a Lowess smoothed version of the data.

To interpret the image data, we analyzed Time within the broader historical context of the 20th century. Our analysis consisted of a close reading of selected issues and articles, chosen based on the following criteria: 1) outlier issues from Fig. 2 defined as those between 1940 and 1990 with > 40% women, 2) issues and articles that were referenced in our secondary sources, and 3) results from EBSCO's Academic Search Complete database, which we used to retrieve all articles in which the word 'woman' or 'women' or 'housewife' was mentioned within our dates of interest.

Our quantitative data coincides closely with the findings of our qualitative analysis. The number of images of women in Time magazine increases during the second world war as women's role in the workforce expanded beyond traditional 'feminine' occupations to fill the gap and to satisfy increased production in the defence industry[3]. In the post-war years, the number images of women decreases as women across North America were instructed by "social engineers, such as psychologists, that they needed to be good wives and mothers in order to fit normally into post-war life"[4]. Our close reading analysis revealed that during these years, notable women who made the news bore the label of 'housewife', regardless of whether or not they were actually a housewife. The number of images of women increases once again as the women's liberation movement began to take shape in the mid-sixties, peaking at the height of the movement in the 1970s. During this time, the women's movement pushed the institutional structure of Time magazine to change from a gender-based caste system to a more equitable work place[5]. In the 1980s, there is a decrease in images, and we believe that this drop in the representation of women is consistent with the analysis in Susan Faludi's well-known book Backlash[6], in which she describes the 1980s as a decade that rejected feminism. The representation of women then starts increasing once again in the 1990s and onward.

Conclusion

We found that a distant reading of the images of faces in Time magazine is consistent with a historical analysis of American socio-political trends and with a close reading of the magazine's content. Specifically, we found that the percentage of female faces peaks during eras when women have been more active in public life, and wanes in eras of backlash against women's rights. This finding is particularly relevant in our contemporary post-literate world in which people absorb culture through images, and spend more time scanning images than reading print content.

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, "Focal Loss for Dense Object Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence (2018): 1-1, https://doi.org/10.1109/TPAMI.2018.2858826;

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal Loss for Dense Object Detection," ArXiv:1708.02002 [Cs] (August 2017), http://arxiv.org/abs/1708.02002.

[2] P Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep Face Recognition," in Proceedings of the British Machine Vision Conference 2015, 41.1-41.12 (Swansea: British Machine Vision Association, 2015), https://doi.org/10.5244/C.29.41; Refik Can Malli, "VGGFace Implementation with Keras Framework. Contribute to Rcmalli/Keras-Vggface Development by Creating an Account on GitHub," GitHub, [2016] 2019, accessed September 28, 2019, https://github.com/rcmalli/keras-vggface.

[3] Ellen Carol DuBois and Lynn Dumenil, Through Women's Eyes: An American History with Documents, 2nd ed. (Boston: Bedford/St. Martin's, 2009), 548-53.

[4] Mona Gleason, Psychology, Schooling, and the Family in Postwar Canada (Toronto: University of Toronto Press, 1999), 52.

[5] Curtis Prendergast and Geoffrey Colvin, The World of Time Inc: The Intimate History of a Changing Enterprise 1960 - 1980 (New York: Atheneum, 1986)

[6] Susan Faludi, Backlash: The Undeclared War Against American Women (Anniversary edition) (New York: Broadway Books, 2006).

Metaphors we read by: Finding metaphorical conceptualizations of reading in web 2.0 book reviews

J. Berenike Herrmann
berenike.herrmann@unibas.ch
University of Basel, Switzerland

Thomas C. Messerli
thomas.messerli@unibas.ch
University of Basel, Switzerland

Keywords: metaphor, social reading, cultural analytics, metaphor identification
Topics: Comparative (2 or more geographical areas), Europe, English, Contemporary, cultural analytics, semantic analysis, Cultural studies, Literary studies

Metaphors we read by: Finding metaphorical conceptualizations of reading in web 2.0 book reviews

Introduction

While interdisciplinary research on metaphor is abundant (Eggs, 2000; Semino & Demjén, 2017; Veale et al., 2016), it is still scarce in Digital Humanities. At the intersection of literary studies, corpus stylistics, and digital humanities, we present an exploratory quantitative metaphor analysis of a corpus of German language lay book reviews. Using a deliberately simple methodological approach that operates on seed words for conceptual sources and targets we investigate how reading experiences of literary texts are metaphorically presented by reviewers.

We explore a corpus of approx. 1.3 mill. book reviews for metaphors used to conceptualize the target domain READING EXPERIENCE.

In line with conceptual metaphor theory, metaphors in language are understood as closely linked to human thought processes and experiences (Lakoff & Johnson, 1980, pp. 4–6; Shutova, 2017). They are mappings from typically more basic experiential source domains (LIFE) to more abstract target domains (READING EXPERIENCE), indicated by indirectly used lexis (the words come, end,and road in "we've come to the end of our road", VUAMC, Steen et al., 2010).

Starting from findings on literature reviews in English (Stockwell, 2009; Nuttall & Harrison, 2018) and on reviews in German (Köhler, 1999), we analyze metaphor patterns in social reading networks, with a particular focus on the mapping READING EXPERIENCE IS MOTION. The main aim at this stage is to draw up a first typology of mappings.

Method and Data

Metaphor Identification

In view of the challenges of reliable automatic metaphor detection (Veale et al., 2016), we apply a deliberately simple rule-based corpus stylistic approach (Deignan & Semino, 2010). A commonly used resource for identification of metaphorical lexical items per source domain is semi-automatic semantic tagging (Demmen et al., 2015). However, in the absence of an out-of-the-box semantic tagger for German, we rely on a 'traditional' onomasiological resource (Dornseiff, 2004). Metaphors are identified by (1) detecting seed words for target domains, (2a-c) detecting source domain seed words in the textual neighborhood of target domain seed words: the metaphor vehicles. Potential metaphors are examined and assigned to a typology of mappings by inspection of KWICs (3).

Step 1. To identify target domain seed words, we compile a list of 'objects of reading experience' (OREs), i.e. noun lemmas that refer to aspects of reading (literary works, such as Buch'book', Geschichte'story', Roman'novel' and parts thereof, such as Ende 'ending' or Spannung'suspense', see Table 1).

Step 2. (2a) Potential source domains are pre-identified by manual MIPVU annotation of small samples of the data (cf. Herrmann et al., in press), and the literature on 'reading' metaphors (e.g. Nuttall & Harrison, 2018). For the present paper, we focus on conceptualizations of reading experiences as MOTION (see Herrmann & Messerli, submitted, for metaphor vehicles from the domain FOOD INTAKE). (2b)The lexical access points to the MOTION domain are provided by a word list extracted from Dornseiff (2004) for the semantic field Fortbewegung (8.3, see Table 2). (2c)To find potential metaphor vehicles that refer to ORE (and not to some other referent), cooccurrences are computed between 'motion' lemmas and ORE, with a window of 10 lemmas around ORE (using raw frequencies, see Table 3). [1]

Lemma	*Tokens*
Buch ('book')	3040,811
Geschichte ('story')	1676,990
Ende ('end/ending')	619,137
Seite ('page/side')	589,534
Charakter ('character')	571,111
Leser ('reader')	559,430
Band ('volume/band')	552,635
Roman ('novel')	512,693
Teil ('part')	505,592
Autorin ('author-F')	502,238
Schreibstil ('writing style')	417,202
Handlung ('plot')	344,847
Spannung ('suspense')	304,937
Autor ('author-M')	280,446
Reihe ('series/row')	267,906
Anfang ('beginning')	266,810
Protagonist ('protagonist')	230,778
Story ('story')	215,364
Thema ('theme/topic')	212,623
Person ('person')	209,701
Kapitel ('chapter')	194,654
Figur ('figure/character')	184,541
Total No. ORE	12,259,980

Table 1: Objects of Reading Experience (ORE) in LoBo

Lemma	No. of Tokens
gehen ('to go/walk')	685,961
ziehen ('to draw pull drag')	124,291
laufen ('to run')	77,822
lebendig ('alive/lively')	51,841
fliegen ('to fly')	45,756
reisen ('to travel')	41,376
stoßen ('to push')	41,358
fahren ('to drive/ride')	38,166
wechseln ('to change')	35,960
fliehen ('to flee')	29,664
...	...
Total MOTION lemmas	1,481,032

Table 2: Ten most frequent potential MOTION seed words in LoBo

Lemma (semantic field: Fortbewegung 'motion')	Freq. within ten words of an ORE
gehen ('to go/walk')	163,320
ziehen ('to draw pull drag')	45,021
fliegen ('to fly')	29,040
lebendig ('alive/lively')	25,743
wechseln ('to change')	14,627
Fahrt ('drive/ride')	14,254
laufen ('to run')	12,613
steigen ('to rise/climb')	11,566
Tempo ('speed')	10,992
starten ('to start')	8,322
...	...
Total Cooccurrences	389,689

Table 3: Cooccurrences. Most frequent MOTION seed words within a window of ten words of an ORE

Step 3. From the resulting frequency list of potential 'motion'-metaphor vehicles (n= 389,689) a sub-section of the most common lemmas is examined by means of KWICs to determine whether potential vehicles were indeed used metaphorically. In a qualitative step, we infer usage patterns from the resulting true metaphor positives (Table 4).

Data

The LoBo corpus (extracted from the social reading platform "Lovelybooks") contains approx. 1.3 mill. German language reviews by 54,000 users, amounting

to 439,923,000 words (Table 4), spread over 15 genres. Each review features a rating (1–5 stars) that refers to a specific book. The corpus is lemmatized and PoS-tagged with TreeTagger (Schmid, 1994), and encoded in CWB (http://cwb.sourceforge.net/).

Type of lemma	Word Count
Target seed words (ORE)	12,259,980
Source seed words (MOTION)	1,481,032
Cooccurrences ORE + MOTION seed word (window=10 w.)	389,689
Total words in corpus	439,923,009

Table 4: Overview of word frequencies of ORE and source domain seed words in LoBo

Analysis

A first result is a list of those lemmas from the semantic field Fortbewegung'motion' that occur frequently within a window of ten words of ORE. While it does not yet allow for conclusive results regarding metaphor use, this list serves as an intermediary step towards identifying a multitude of MOTION metaphors for subsequent analysis establishing a typology of mapping patterns.

The analysis of KWICs shows that certain manners of motion are particularly frequent. Notable are the motions of walking, flying, and driving/riding, realized with the lemmas gehen'to go', fliegen 'to fly', and Fahrt'ride/drive'. Notable is variance of 'speed', with fast motion(Fahrt, fliegen), and slower motion (gehen).

Another important observation is about agency within the metaphorical scenario. Readers position themselves mainly as (a) observers who see how the plot moves along; (b) agents who actively 'walk' and 'fly' through the story (or a book's pages); (c) patients being put in motion by the book; and (d) companions who travel along with an ORE (see Table 5). Findings demonstrate the complexity of reading that cannot be restricted to passive reception or hedonistic consumption (cf. Rebora et al., 2019).

1) an ORE is the moving agent	Die <**Seiten fliegen**> nur so dahin. 'The pages are simply flying.'
	Category 1 consists of conceptualizations of reading in which the reader is not given an explicit role, but can be inferred to be an observer. An ORE is the central element and is described as moving through space.
2) the reader is the moving agent	Geschrieben ist das Buch wunderbar fließend , man <**fliegt förmlich durch die Seiten**> 'The book has a wonderful flow (lit. is written wonderfully flowingly), you are positively flying through the pages'
	Category 2 positions the reader as someone who moves through space and an ORE as a different element of that MOTION (e.g. a space through which the reader moves)
3) an ORE as agent moves the reader or another ORE as a patient	Der Spannungsbogen , der sich bis zum Ende hält , hat mich geradezu durch das <**Buch fliegen**> lassen und ich konnte es kaum aus der Hand legen . 'The arc of suspense that spans until the end has made me positively fly through the book and I almost couldn't put it aside.'
	Category 3 positions an ORE as the instigator of movement of another ORE or the reader.
4) an ORE and the reader move together	Der Leser <**pilgert quasi in der Handlung**> mit... 'The reader sort of goes on a pilgrimage with the plot...'
	Category 4 conceptualizes reading as a joint motion between reader and ORE.

Table 5: Categories of mappings READING IS MOTION in LoBo

In all, our study offers a first typology of metaphorical MOTION-mappings in digital shared reading, as well as evidence of the productivity of MOTION as a source domain for READING in German lay reviews (cf. Nuttall & Harrison, 2019, for English reviews). Extending this exploratory phase into statistical analysis, we plan variance analysis with factors as reader's evaluation (star ratings) and book genre (e.g., middle brow vs. popular). Methodologically, we plan to improve precision of metaphor detection, e.g. by including semantic information from resources such as GermaNet, but also through active learning. Generally, further examination of metaphors will allow valuable insight into underlying conceptual and value systems in reader reviews.

References

Demmen, J., Semino, E., Demjén, Z., Koller, V., Hardie, A., Rayson, P., & Payne, S. (2015). A computer-assisted study of the use of Violence metaphors for cancer and end of life by patients, family carers and health professionals. International Journal of Corpus Linguistics, 20(2), 205–231.

Dornseiff, F. (2004). Der deutsche Wortschatz nach Sachgruppen, Mit einer lexikographisch-historischen Einführung und einer ausführlichen Bibliographie zur Lexikographie und Onomasiologie (8th, fully revised version with alphabetical register, reprint 2010).

Eggs, E. (2000). Metapher, in: Gert Ueding (Hg.), Historisches Wörterbuch der Rhetorik, Bd. 5, Tübingen, S. 1109-1183.

Genette, G. (1989). Paratexte: Das Buch vom Beiwerk des Buches. (D. Honig, transl.). Frankfurt am Main: Campus.

Herrmann, J.B. & Messerli, Th. (submitted). ...hungere schon nach dem nächsten Band. Eine Untersuchung von Metaphern für Leseerfahrungen in Web 2.0 Literaturrezensionen. 7. Jahrestagung «Digital Humanities im deutschsprachigen Raum» (DhD), Paderborn, D.

Herrmann, J. B., Woll, K., & Dorst, A. G. (in press). Linguistic metaphor identification in German, in: S. Nacey, A.G. Dorst, T. Krennmayr, & W. G. Reijnierse (Hg.), MIPVU in Multiple Languages, Amsterdam: John Benjamins.

Köhler, M. (1999). Wertung in der Literaturkritik: Bewertungskriterien und sprachliche Ausdrucksmöglichkeiten des Bewertens in journalistischen Rezensionen zeitgenössischer Literatur (PhD Dissertation, Universitätsbibliothek der Universität Würzburg). Retrieved from https://opus.bibliothek.uni-wuerzburg.de/frontdoor/index/index/docId/784

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. Chicago, IL: University of Chicago Press.

Nuttall, L., & Harrison, C. (2018). Wolfing down the Twilight series: Metaphors for reading in online reviews, in: H. Ringrow & S. Pihlaja (Hg.), Contemporary Media Stylistics. New York: Bloomsbury Academic.

Rebora, S., Piroska L., & Kuijpers, M. (2018). Reader Experience Labeling Automatized: Text Similarity Classification of User-Generated Book Reviews. In EADH2018. Galway: EADH. https://eadh2018.exordo.com/programme/presentation/90.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees.

Semino, E., & Demjén, Z. (2017). The Routledge Handbook of Metaphor and Language. London, New York: Routledge.

Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., & Krennmayr, T. (2010). VU Amsterdam Metaphor Corpus. Retrieved from http://www.ota.ox.ac.uk/headers/2541.xml

Stockwell, P. (2009). Texture: a Cognitive Aesthetics of Reading. Edinburgh: Edinburgh University Press.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. Amsterdam: John Benjamins.

[1] Our aim here is not the identification of phraseological units by significance tests against chance distribution (Mutual Information, DICE, and log-likelihood). Rather, raw frequencies allow us to define a window of reference for metaphor vehicles.

Prototyping the SF Nexus: Collaborative Models for Digitizing and Curating Speculative Fiction Collections as Data

Henry Alexander Wermer-Colan
alex.wermer-colan@temple.edu
Temple University, United States of America

Rikk Mulligan
rikk@cmu.edu
Carnegie Mellon University, United States of America

Keywords: Science Fiction, Digital Collections, Cultural Analytics
Topics: Europe, English, North America, 19th Century, 20th Century, Contemporary, cultural analytics, digital libraries creation, management, and analysis, Book and print history, Literary studies

This paper overviews the SF Nexus prototype, the developmental stage of a research project to digitize and curate available works of Anglo-American speculative fiction. This resource enables access to texts in the public domain and under copyright, from magazines to mass-market novels. We overview challenges confronting scholars of SF book history, including restricted access to and the uncatalogued materials in special collections, copyright barriers to sharing digital texts, and the lack of comprehensive indices to SF texts and their publication records. The presentation will present models for multi-institutional, collaborative mass digitization efforts to ingest into HathiTrust, share across research centers and libraries, and curate on the web with user-friendly, embeddable tools for the cultural analytics of books as image and text data. We conclude by proposing our coalition as a collaborative model for similar digitization and curation projects requiring standardized policies, legal agreements, and data curation workflows.

Sociocultural trend signatures in minimal persistence and past novelty

Kristoffer Nielbo
kln@cas.au.dk
Center for Humanities Computing Aarhus, Aarhus University, Denmark DATALAB, Aarhus University, Denmark

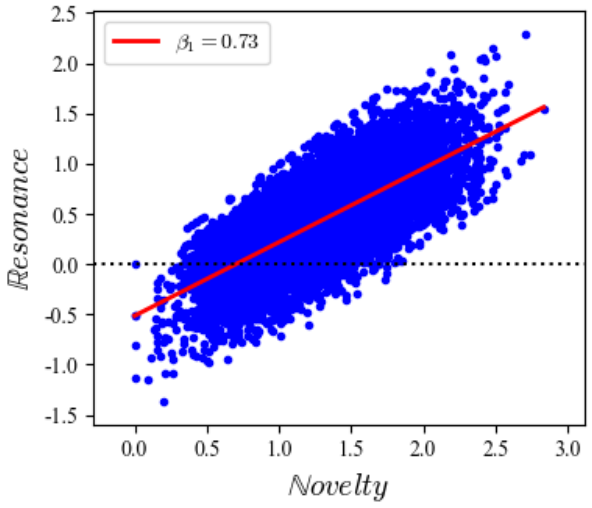
Peter Bjerregaard Vahlstrup
Center for Humanities Computing Aarhus, Aarhus University, Denmark DATALAB, Aarhus University, Denmark

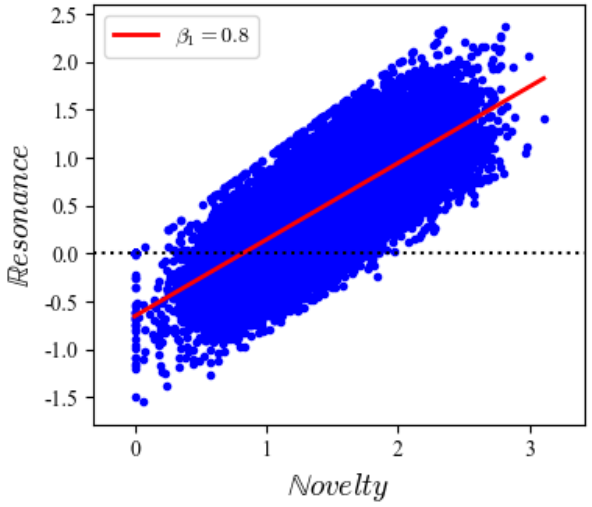
Jianbo Gao
Center for Geodata and Analysis, Faculty of Geographical Science, Beijing Normal University, China Institute of Automation, Chinese Academy of Sciences, China

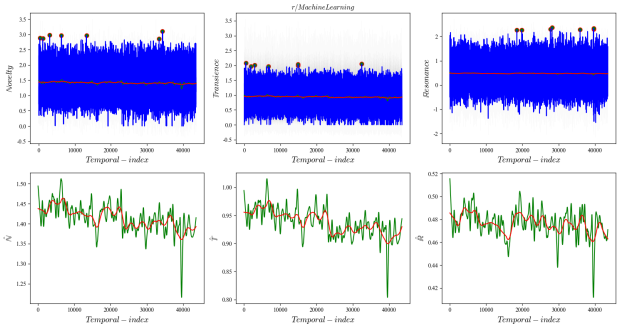
Anja Bechmann
DATALAB, Aarhus University, Denmark

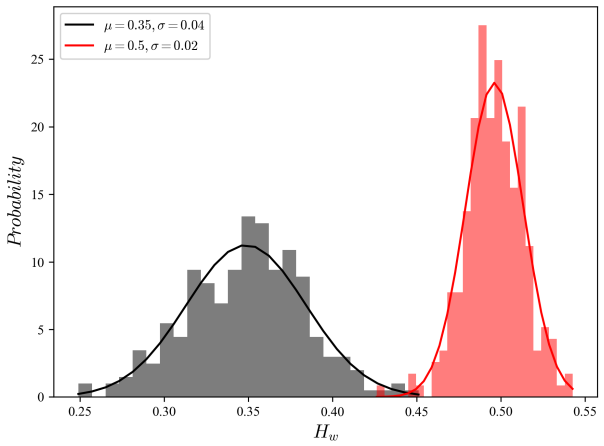
Keywords: trend estimation, dynamical systems, fractal analysis, information theory, cultural analytics
Topics: Asia, Europe, English, North America, Contemporary, cultural analytics, information retrieval and querying algorithms and methods, Humanities computing, Media studies

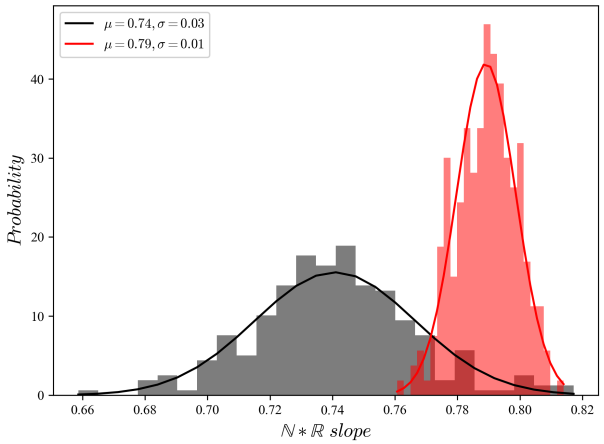
Sociocultural trends from social media platforms such as Twitter or Instagram have become an important part of knowledge discovery. The `trend' construct is however ambiguous and its estimation from unstructured sociocultural data complicated by several methodological issues. This paper presents an approach to trend estimation that combines (`intersects') domain knowledge of social media with advances in information theory and dynamical systems. In particular, we show how *trend reservoirs* (i.e., signals that display trend potential) can be identified by their relationship between novel and resonant behavior, and their minimal persistence.This approach contrasts with trend estimation that relies on linear or polynomial techniques to study point-like novelty behavior in social media, and it completes approaches that rely on smooth functions of time.











Mind the Gap: Bridging Distant and Close Reading across Heterogeneous Text Collections

Glenn Roe
glenn.roe@sorbonne-universite.fr
Sorbonne University, France

Clovis Gladstone
clovisgladstone@uchicago.edu
University of Chicago, USA

Mark Olsen
markymaypo57@gmail.com
University of Chicago, USA

Robert Morrissey
University of Chicago, USA

Keywords: intertextuality, distant reading, text similarity, big data, machine learning
Topics: Europe, English, North America, 18th Century, information retrieval and querying algorithms and methods, text mining and analysis, Book and print history, Literary studies

This paper describes the development of a new model of scholarly interaction that allows users to bridge the gap between distant and close reading approaches when conducting research on large, heterogeneous digital text collections. Using this model, which we call the 'intertextual bridge', our aim is to demonstrate that the conceptual relationships among texts discovered by text-mining algorithms can fruitfully guide close reading in dialectical interaction with distant reading. Fundamentally, we are contending that the core of scholarly reading in the digital age should be the discovery and navigation of intertextual relationships. The 'intertextual bridge' model will thus be a powerful hermeneutical device allowing users to navigate between individual texts and larger corpora that are related through shared themes, ideas, and passages. For this paper, we are focusing on the French Revolutionary period, and we test this model by applying it to several extensive and diverse 18th-century French collections.

"Exalting the Cult of Gentlemanly Amateurism"[1]: Improving Computational Analysis with Humanities Narrative-Building Methodologies

M. H. Beals
Loughborough University, United Kingdom

Emily Bell

Loughborough University, United Kingdom

Julianne Nyhan
University College London, United Kingdom

Tessa Hauswedell
University College London, United Kingdom

Keywords: digitisation, newspapers, metadata, linked open data, history
Topics: Global, English, 19th Century, 20th Century, Contemporary, digitization (2D & 3D), metadata standards, systems, and methods, History, Library & information science

The nineteenth-century newspaper was a messy object, filled with an ever-changing mix of material-literary, factual and the suspiciously plausible-in an innumerable number of amorphous layouts. Working with digitised newspapers is no different. Each database contains a theoretically-standardised collection of data, metadata, and images, but the precise nature and nuance of this data is often occluded by the automatic processes that encoded it. Moreover, no true universal standard has been implemented to facilitate cross-database analysis, encouraging digital research to remain within existing institutional or commercial silos.[2] Where common standards have been asserted, such as the minimum standards for Europeana or Chronicling America, they have been standardised at only a very low resolution, with significant variance in the range and interpretation of the metadata within their direct collaborations as well as by independent programmes following their example.[3] These irregularities make the data highly vulnerable to misinterpretation by both end users and also those updating the collections in the future.

In order to better explore global exchanges (for example, scissors-and-paste journalism) in the nineteenth-century press, Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840-1914 attempted to integrate and make interoperable the metadata used to store digitised newspapers in a variety of linguistic and institutional contexts. This paper will demonstrate how we excavated institutional decision-making from a variety of sources in order to understand the archaeology of digitised newspaper metadata, its vocabulary and structures, and how they related to the conceptions of the newspaper object by both modern end-users and the original nineteenth-century producers. It will explore how computational thinking and data management processes can be combined with discussions of the historical evolution of the newspaper to restore and integrate a narrative that is generally lost in the creation of digital archives: how the strategies and decision-making processes that shaped the composition and structure of the data have, and will continue to, impact user experience and the conclusions drawn from these materials.[4]

The Ontologies work package of Oceanic Exchanges had a simple remit: to catalogue and map the metadata terminology used by newspaper databases to one another and to an internal ontology, to support research into reprinting. Because complete documentation was not available for any of our collections, we retro-engineered the implementation of these vocabularies, beginning with document type definitions (DTDs) and schema specifications, complementing them with internal and public documentation on the cataloguing standards used. Some cases also required us to rely upon grey literature-discussions by users about how to manipulate the data- and direct examination of records. Finally, building upon previous research by team members and new interviews, we were able to develop a longitudinal understanding of how the data has been augmented or repackaged by institutions over the past twenty years.

Although most of the databases used variants of the METS/ALTO standard, these were not implemented in a way that would allow for simple equivalencies. The variance in terminology, and in the interpretation of the correct range of inputs for a given field, arose from the use of a hodgepodge of different vocabularies, including variants of Dublin Core, METS/ALTO, MPEG-21, PREMIS, as well as other bespoke or proprietary taxonomies. Overlapping and ambiguous vocabularies were also structured inconsistently, with some combining data at the article, page or issue level and others separating the metadata and content for these elements into multiple files. Our initial attempts to account for both internal structures and field equivalencies across these databases made the level of irregularity strikingly clear.

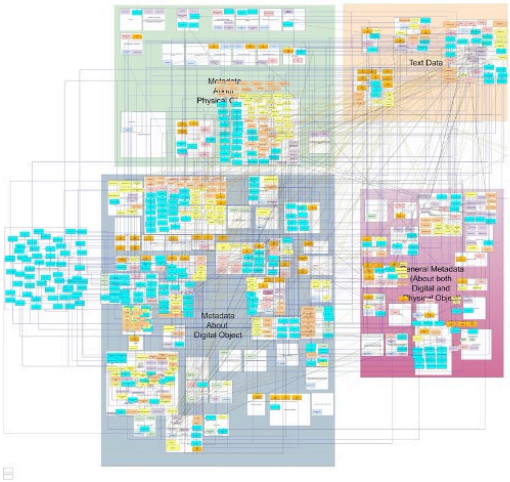


Figure 1: Map of all metadata fields from our samples (each one represented by a different colour), with connecting lines showing the internal hierarchy of each, broken down by metadata of physical object, digital object, metadata pertaining to both, and text data. Unmapped blue boxes represent an overflow of repetitive administrative technical metadata.

Moreover, the interpretation and implementation of these fields was inconsistent within collections owing to the turnover of staff during the digitisation process as well as the long history of metadata being drawn from existing library catalogues. Such layering is particularly evident in the metadata associated with Trove, the National Library of Australia's collections, which includes end-user annotations, categorisations and text corrections-layers which are valuable to humanities researchers but which remain in unintegrated grey literature and derived data for the other collections. The level of publically-available documentation about how to interpret both authorised and user-generated fields varied widely, and interviews and internal documents made it clear that consistent implementation of guidelines was unlikely across time. Working with these collections, therefore, requires a creative and flexible interpretation of these standards and an understanding of the history and character of the specific digital files.

After working with such disparate source materials, we concluded that the narrative of creation, archiving and digitisation might be most robustly and sustainably documented through a decentralised and layered medium, namely Linked Open Data. This decision is not without controversy. First, although the scale of periodical material makes it particularly tempting for large-scale analysis, the majority of newspaper metadata is in XML format, which presents specific challenges for semantic data modelling.[5] More philosophically, the possibilities and problematics of the semantic web have been theorised since the term was coined in 2001;[6] in particular, the importance of making and sustaining connections to humanistic forms of knowledge representation has been regularly emphasised.[7] Oldman, Doerr and Gradmann highlight the possibility of linked data combining "digital infrastructure, computer reasoning, interpretation, and digital collaboration", but warn of leaving a "mechanical meaningless shell"[8] if the endeavour is seen as an end in itself, or as a purely scientific exercise.[9] Indeed, linguistics and literary scholars have raised numerous concerns about tool-driven research questions and banal quantification for computation's sake.[10] Likewise, Berry and Fagerjord have claimed that linked data involves a fragmentation that "privileges knowledge divided into non-narrative shards of information",[11] seemingly putting it in direct opposition to the idea of reclaiming lost narratives of creation and use.

This paper will, therefore, explore the implications of competing claims surrounding the value of Linked Open Data within the specific domain of digitised periodicals, particularly when working with enriched metadata and data roundtripping (the process of integrating derived data back into the original collections), and demonstrate how combining institutional histories, interviews, metadata and historical narrative detail in a decentralised and layered structure can restore lost narratives and data provenance in a sustainable way that is intelligible across disciplines and at multiple resolutions-whether focusing on the textual content of the issue, the technical details surrounding digitisation, or the computational representation of the physical layout and materials.

[1] Noam Chomsky, Language and Mind (Cambridge: Cambridge University Press, 2006): 19.

[2] Developments in this area have been led by the Europeana Data Quality Committee, which has developed a Metadata Quality Assurance Framework for Europeana (see http://144.76.218.178/europeana-qa/), but engagement with this is limited by the project partners' existing metadata practices. Other cultural heritage metadata ontologies such as the CIDOC Conceptual Reference Model (http://www.cidoc-crm.org/) and FRBRoo (https://www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf) do not include newspapers.

[3] See, for example, the guidelines provided by three key digisters, which generally conform to agreed METS/ALTO standards, but vary (and allow for variance) in metadata range and interpretation. "Technical Guidelines for 2018 Awards", Library of Congress, 18 December 2018, https://www.loc.gov/ndnp/

guidelines/archive/guidelines1819.html. Accessed 10 October 2019; "Europeana Data Model", Europeana Pro, 2017, https://pro.europeana.eu/resources/standardization-tools/edm-documentation. Accessed 10 October 2019; Although internal documentation of METS/ALTO guidelines is not publicly available (it has been reviewed by the authors), the guidelines for the public API (with simplified field names) provide similar insights. "API version 1 technical guide", National Library of Australia, 2018, https://help.nla.gov.au/trove/building-with-trove/api-version-1-technical-guide. Accessed 10 October 2019

[4] See Bonnie Mak, "Archaeology of a Digitization", Journal of the American Society for Information Science and Technology 65.8 (2014): 1515-26 and Paul Fyfe, "An Archaeology of Victorian Newspapers," Victorian Periodicals Review 49.4 (2016): 546-77.

[5] See Fabio Ciotti and Francesca Tomasi, "Formal Ontologies, Linked Data and TEI", Journal of the Text Encoding Initiative 9 (2016-2017): http://journals.openedition.org/jtei/1480. Accessed 10 October 2019.

[6] See Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", Scientific American, (2001).

[7] See Dominic Oldman, Martin Doerr and Stefan Gradmann, "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge" in Schreibmann, Siemens and Unsworth, eds. A New Companion to Digital Humanities (Oxford: Wiley-Blackwell, 2016): 251-73.

[8] Ibid. 252.

[9] Sean Bechhofer et al, "Why linked data is not enough for scientists", Future Generation Computer Systems 29.2 (2013): 599-611.

[10] Chomsky, Language and Mind: 19; Nan Z. Da, "The Computational Case Against Computational Literary Studies," Critical Inquiry 45.3 (2019): 601-39.

[11] David M. Berry and Anders Fagerjord, Digital Humanities: Knowledge and Critique in a Digital Age (Cambridge: Polity, 2017): 77.

Trans-Hispanic Networks of Feminist Solidarity: The Rise and Spread of #8M

Vanessa Ceia
vanessa.ceia@mcgill.ca
McGill University, Canada

Rhian Lewis
rhian.lewis@mail.mcgill.ca
McGill University, Canada

Keywords: Feminism, Twitter, #8M, Hispanic Feminisms, Digital Activism
Topics: South America, Comparative (2 or more geographical areas), Europe, English, Contemporary, digital activism and advocacy, social media analysis and methods, Cultural studies, Feminist studies

On March 8 2019, Spanish-speaking Twitter communities erupted with a polyvocal outcry against gender violence, while simultaneous offline protests took place in cities across the globe. This paper presents the results of a multi-modal spatial and thematic analysis of regional appropriations of the hashtag #8M (8 March) to illustrate the carrefours/intersections of trans-Hispanic networks of feminist solidarity online. Through spatial, network, and word frequency analyses of #8M, we examine the hashtag's intersections with existing transnational and regionally-specific feminist Twitter dialogues, including the #NiUnaMenos movement launched to combat femicide in Argentina and the hashtag #Cuéntalo used to protest the lenient sentencing of a group of men who raped a woman in Pamplona, Spain. By combining quantitative and geospatial analyses using Twitter Archiving Google Sheets, Voyant-Tools, and Carto with targeted close readings of tweets containing #8M, this paper traces the regional variations in a large, multinational online Spanish-language conversation about gender violence.

Tribesourcing Southwest Films: Counter-Narrations and Speaking Back to Colonial Rhetoric

Melissa Dollman
msdollman@unc.edu
University of North Carolina at Chapel Hill, United States of America

Jennifer Jenkins
jenkinsj@email.arizona.edu
University of Arizona, United States of America

Rhiannon Sorrell
rsorrell@dinecollege.edu
Diné College, United States of America

Keywords: counter-narration, indigenous studies, repatriation, decolonization, crowdsourcing
Topics: English, North America, 20th Century, Contemporary, crowdsourcing, public humanities collaborations and methods, Film and cinema arts studies, First nations and indigenous studies

The *Tribesourcing Southwest Film Project* seeks to "tribesource" counter-narrations for educational films about the Native peoples of the Southwestern U.S. These films originate from the American Indian Film Gallery, a collection awarded to the University of Arizona in 2011.

They were made in the mid-20th century for the K-12 educational and television markets and reflect mainstream cultural attitudes of the day. Often the male "voice of God" narration pronounces meaning that is inaccurate or disrespectful, but the visual narratives are for the most part quite remarkable. At this historical distance, many of these films have come to be understood by both Native community insiders and outside scholars as documentation of cultural practices and lifeways-and, indeed, languages-that are receding as practitioners and speakers pass on. This project seeks to rebalance the historical record through digital intervention, intentionally shifting emphasis from external perceptions of Native peoples to the voices, knowledge, and languages of the peoples represented in the films by participatory recording of new narrations for the films. Native narrators record new narrations for the films, actively decolonizing this collection and performing information redress through the merger of vintage visuals and new audio. We propose to report on work-in-progress and future plans.

"Tribesourcing" obviously plays off the term "crowdsourcing," but intentionally places historical materials with the exact peoples they represent so they may address the untold or suppressed story. Furthermore, Native narrators are compensated for their contributions, and identified: aspects that temper what might otherwise be perceived as another example of soliciting "free" and "invisible" labor from communities of color. Films in this project are streamed through the interface of our chosen content management system, Mukurtu, beside alternate narrations from within the culture in English and in Native languages. This method allows for an unmediated approach to identification of people, places, practices, vocabulary and stories that might otherwise be lost, as well as providing a rich, community-based metadata record for each film. Taking a small step toward cultural repatriation of content, tribesourcing as a methodology is guided by the Protocols for American Indian Archival Materials (2006) and the groundbreaking work of Margaret Kovach (2010) and Linda Tuhiwai Smith (2012).

The process of identifying narrators is best handled within the communities.Thus, paid Tribal Narration Coordinators perform the essential function of recruiting and working with narrators on-site. Rhiannon Sorrell serves in this position at Diné College in Tsaile, Arizona. Having done the very first counter-narration as proof-of-concept for the Tribesourcing project, Sorrell is well-equipped to explain the project to potential narrators and oversee recordings and payments for narrators. She will give her perspective in our presentation.

From its inception, this project has existed at the crossroads of indigenous studies, film and media studies, and public digital humanities. The content resides with our Native partners; dissemination and access occur through this digital humanities project that we steward. Close collaboration with Native Nations helps to ensure cultural competency in information management, and Mukurtu's incorporation of a digital toolset (Google maps, HTML5 video and audio players, linked data capabilities, and integration with third-party streaming platforms) afford us the means of reaching a broad user base in Indian Country, where geographical distances often inhibit access to primary materials. The new narrations, metadata, dictionary, and Traditional Knowledge Labels, when placed literally alongside films themselves, help to subvert the films' original narrative conventions, and provoke the viewer to question the visual and spoken rhetoric (Posner, 2016). As a film preservation project, we also acknowledge "'precious fragments' of indigenous knowledge are increasingly held captive in obsolete audio-visual media formats" (Lawson, 2017). Therefore, issues of data migration and storage are integral to the project as well. Digitalized originals and access copies are distributed geographically, and the original films are housed at the Library of Congress.

The undertaking of positioning the *Tribesourcing Southwest Film Project* as an interactive, multimedia, multiethnic, and polyvocal resource raises both workaday and theoretical issues, among them the need for Native language presence in the archive as a whole. Informed, intentional practice is essential. Thus, every stage of the project requires consultation and collaboration with our partners in Native Nations. This project seeks to contribute to ongoing efforts to decolonize the archive and restore voice and narrative sovereignty to the people who appear in these films-

as agents of their own information rather than subjects of a governmental or corporate agenda. Participation, Equity, and Inclusion are central to this project. This project is supported by a grant from the (U.S.) National Endowment for the Humanities.

References

Kovach, Margaret. (2010). Indigenous Methodologies: Characteristics, Conversations, and Contexts. University of Toronto Press.

Lawson, Gerry. (2017). "Indigitization: It Takes a Community." Presentation, Museum of Modern Art, New York, NY. http:// www.digitalassetsymposium.com /2017/06/26/ indigitization-it-takes-a-community/

Posner, Miriam. (2016). "How Is A Digital Project Like a Film?" The Arclight Guidebook to Media History and the Digital Humanities, edited by Charles R. Acland and Eric Hoyt, pp. 184- 194.

Protocols for American Indian Archival Materials. (2006).

Smith, Linda Tuhiwai. (2012). Decolonizing Methodologies: Research and Indigenous People. London: Zed Books.

Traditional Knowledge (TK) Labels. http://localcontexts.org/tk-labels/

Situated Materialities, Object Agency and Resistant Bodies in Subalternized Craft: Neoliberal Market/Tourist Logics and Questions around what/why of the "decolonializing" of digital humanities

Radhika Gajjala
radhik@bgsu.edu
Bowling Green State University, United States of America

Rhiannon Bettivia
University of Illinois, Illinois, USA

Taylar Stagnar
Bowling Green State University, United States of America

Rustina Untari
Soegijapranata Catholic University, Semarang, Indonesia

Keywords: Weaving, archiving, metadata, memory, lost objects
Topics: Asia, Comparative (2 or more geographical areas), English, North America, Contemporary, cultural analytics, ethnographic analysis, Feminist studies, First nations and indigenous studies

This paper draws on ethnographic work in various indigenous and international contexts of digital archiving and on research in subaltern craft communities to think through the ontic/epistemic imaginaries (Verran, 2001) and the issues surrounding the inclusion of subaltern craft communities in digital global space either through museumizing or entrepreneurial promises of reaching global markets. In doing so, we raise questions regarding the scalar politics of "authenticity" from local to supranational contexts while problematizing notions of "untouched" cultural products.

Locating Biographies of Canadian and Native American Women: Nationality, Race, and Mobility in North America

Alison Booth
ab6j@virginia.edu
University of Virginia, United States of America

Keywords: Women, Canadian, Native American, XML, GIS
Topics: Europe, English, North America, 19th Century, 20th Century, digital biography, personography, and prosopography, spatial & spatio-temporal analysis, modeling and visualization, Feminist studies, Literary studies

As part of the next phase of Collective Biographies of Women (CBW), this paper focuses on E. Pauline Johnson as an example of representation of Canadian and Native American women in this database of 1272 books, 8372 persons, and 14,185 chapter-

length biographies. The talk explains the team's digital methods for visualizing spatial data in collective biographies of individuals and cohorts. Interpretation of the historical and ideological contexts, the paper shows, is a necessary framework for reading the significance of spatial data, life paths, and the typologies such as nationality assigned by the printed biographies or the research database. Movement in space is structured by power and historical change: shaped by restrictions on women, people of color, events such as war, shifting national boundaries, and changing means of travel.

Framing the experience: a study of the history of interfaces to digital humanities projects

Claire Warwick
c.l.h.warwick@durham.ac.uk
Durham University, United Kingdom

Keywords: interface design, user studies, history of DH
Topics: Comparative (2 or more geographical areas), English, 20th Century, Contemporary, digital archiving, Interface design, development, and analysis, Humanities computing, Library & information science

Drucker (2013) argues that, although the goals of Human Computer Interface research are to render the interface invisible and facilitate access to digital content, interfaces themselves should be legitimate objects of study. Yet little attention has been paid to this aspect of digital humanities resource design. The following proposal therefore reports on a study of interfaces to long-lived DH resources to determine what information we may gain from them about the history of DH project development.

The Semantics of Structure in Large Historical Corpora

Marijn Koolen
marijn.koolen@di.huc.knaw.nl
KNAW Humanities Cluster

Rik Hoekstra
rik.hoekstra@di.huc.knaw.nl
KNAW Humanities Cluster

Keywords: structure extraction, corpus accessibility
Topics: Europe, English, 15th-17th Century, 18th Century, concordancing and indexing, text mining and analysis, History, Library & information science

Structuring large historical corpora that are too big to be processed manually can take two approaches. The first is an inductive method extracting implicit entities and meaning from textual (and sometimes visual) content. With the help of AI or manually compiled (existing) lists of entities, the entities are converted into information. The second, that Colavizza (2019) calls referential information systems, takes existing reference systems (like archival indexes) and uses them to contextualize individual documents. Both methods are used to turn corpora into computer accessible information systems. Ideally a more complete information system would result from combining both approaches, but in practice they are hard to bridge because of a number of different problems. This paper presents an approach that addresses those problems and combines inductive methods of automated text analysis and information extraction techniques with knowledge of the referential information systems to add rich semantic layers of information to large historical corpora.

Making large historical corpora accessible for research usually involves a pipeline of processing steps, ranging from text recognition to entity and event spotting, disambiguation, identification and ideally contextualization (Meroño-Peñuela et al. 2015). In many projects much effort is spent on producing a close-to-perfect text by transcribing, or by a mixed procedure of automatic transcription by Optical Character Recognition (OCR) or Handwritten Text Recognition (HTR) and manual correction of the results, as many of the later elements in the pipeline require high-quality text to work well. There are ways to partially solve OCR or HTR (Handwritten Text Recognition) errors automatically through post-correction (see e.g. Reynaert 2014, Reynaert 2016), or to use word embeddings to overcome matching problems, (e.g. Egense 2017). The most important limitation of this approach is that full-text alone is not enough to make a corpus available for research that is not primarily directed at the text but rather at its information (Hoekstra and Koolen 2018, Upward 2018). Extracting and contextualizing information has many issues such as OCR and HTR errors that make it

difficult to use standard Natural Language Processing (NLP) tools like Named Entity Recognition (NER), topic modelling, Part Of Speech (POS) tagging and sentiment analysis, which has been common knowledge for a long time (Lopresti 2008, Traub et al. 2015, Mutuvi et al. 2018, Hill & Hengchen 2019, van Strien et al. 2020). However, solutions for such issues are scarce and badly documented, as argued by, amongst others, Piersma and Ribbens (2013), van Eijnatten et al. (2013) and Leemans et al. (2017).

Many archives and libraries have experimented with giving access to their collections by means of their digitized inventories and some have gone a step further, using existing indexes of serial collections (Jeurgens 2016, Colavizza 2019, Head 2003). But these archival referential systems are too coarse for access beyond the document level. However, the existing scholarly apparatus consists of many more reference systems and tools that can be put to good use. Centuries of dealing with these complications have led to a number of convenient and often-employed structures that are part of the printed culture but are often ignored in the translation to digital access (Upward 2018, Opitz 2018).

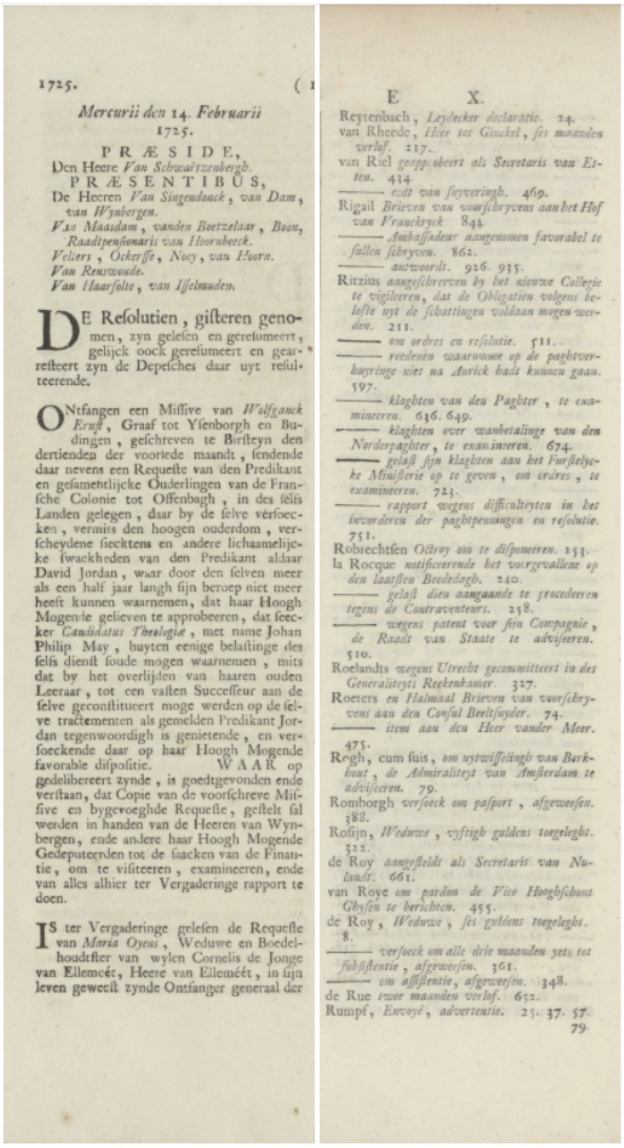
Exploiting Referential Information Systems and Repetitive Phrases

Instead of trying to find latent semantic structures through full-text analysis, these explicit structures allow for finding intended semantic information that is likely not available in another form. Remarkably, many digitization programmes take no advantage of these structures and sometimes do not even digitize them, extracting only the main textual body as plain text.

A concrete and relatively simple example, the Resolutions of the Dutch States General is a collection of all resolutions (decisions) of the Dutch Republic from 1576 until 1796 and contains around 440,000 pages. Roughly half of the pages, up to 1703, are handwritten. From 1703 onwards, the States General printed yearly editions for easier access to previous resolutions. The States General met six days a week and kept a list of who was present on what date, followed by a summary of each resolution. The left side of Figure 1 shows one column from the printed edition of 1725 with the meeting date, attendants list and some of the resolutions. On the right side is a column from an index page with indexed terms and page references.

Typical steps in digitizing such a large resource are the identification of columns on the page and OCR of the text to make it full-text searchable. Most typically NER would be applied to automatically identify named entities, even if its success depends highly on the OCR and HTR quality. With an estimated 1.5 to 2 million mentions of person names and close to half a million geographic names and organisations, NER would result in a huge list of names, with many false positives (incorrect names) and false negatives (unrecognised names). One issue is that NER tools tend to use uppercase initial letters as a signal that a word is a name, but in early modern texts, uppercase was used for many nouns as well. The most frequently recognized person names then tend to be terms that were often mentioned, like Money, Passport and Meeting. The main problem is that there is no way to predict which names are correct and which are not so users cannot easily filter out the obvious errors from the hundreds of thousands of distinct names. Using reference tools with lists of known historical persons for identification has the limitation that they cover only a small selection of the most famous names, which are not necessarily the most relevant names.

We argue that projects should focus more energy on extracting and operationalizing the existing structure of such corpora. In the case of the resolutions, there is a lot of value in recognizing the meeting dates, attendants lists and index terms.



1725. (1

Mercurii den 14. Februarii 1725.

PRÆSIDE,
Den Heere *Van Schwartzenbergh.*
PRÆSENTIBUS,
De Heeren *Van Singendouck, van Dam, van Wynbergen.*
Van Maasdam, vanden Boetzelaar, Boon, Raadtpenſionaris van Hoornbeeck.
Velters, Ockerſſe, Nocy, van Hoorn.
Van Renswoude.
Van Haarſolte, van Iſſelmuden.

DE Reſolutien, giſteren genomen, zyn geleſen en gereſumeert, gelijck oock gereſumeert en gearreſteert zyn de Depeſches daar uyt reſulteerende.

ONtfangen een Miſſive van *Wolfganck Ernſt*, Graaf tot Yſenborgh en Budingen, geſchreven te Birſteyn den dertiendẽ der voorlede maandt, ſendende daar nevens een Requeſte van den Predikant en geſamentlijcke Ouderlingen van de Franſche Colonie tot Offenbagh, in des ſelfs Landen gelegen, daar by de ſelve verſoecken, vermits den hoogen ouderdom, verſcheydene ſiecktens en andere lichaamelijcke ſwackheden van den Predikant aldaar David Jordan, waar door den ſelven meer als een half jaar langh ſijn beroep niet meer heeft kunnen waarnemen, dat haar Hoogh Mogende gelieven te approbeeren, dat ſeecker *Candidatus Theologiæ*, met name Johan Philip May, buyten eenige belaſtinge des ſelfs dienſt ſoude mogen waarnemen, mits dat by het overlijden van haaren ouden Leeraar, tot een vaſten Succeſſeur aan de ſelve geconſtitueert moge werden op de ſelve tractementen als gemelden Predikant Jordan tegenwoordigh is genietende, en verſoeckende daar op haar Hoogh Mogende favorable diſpoſitie. WAAR op gedelibereert zynde, is goedtgevonden ende verſtaan, dat Copie van de voorſchreve Miſſive en bygevoeghde Requeſte, geſtelt ſal werden in handen van de Heeren van Wynbergen, ende andere haar Hoogh Mogende Gedeputeerden tot de ſaacken van de Finantie, om te viſiteeren, examineeren, ende van alles alhier ter Vergaderinge rapport te doen.

IS ter Vergaderinge geleſen de Requeſte van *Maria Oyens*, Weduwe en Boedelhoudſter van wylen Cornelis de Jonge van Ellemeét, Heere van Ellemeét, in ſijn leven geweeſt zynde Ontfanger generaal der

E X.

Reytenbach, *Leydecker declaratie.* 24.
van Rheede, *Heer tot Gruckel, ſes maanden verlof.* 217.
van Riel *geapprobeert als Secretaris van Eten.* 414
——— *eedt van ſtyveringh.* 469.
Rigail *Brieven van voorſchryvens aan het Hof van Vranckryck* 844
——— *Ambaſſadeur aangenomen favorabel te ſullen ſchryven.* 862.
——— *antwoordt.* 926. 935.
Ritzius *aangeſchreeven by het nieuwe Collegie te vigileeren, dat de Obligatien volgens beloſte uyt de ſchattingen voldaan mogen werden.* 211.
——— *ons ordres en reſolutie.* 511.
——— *redenen waarom op de paghtverhuyringe niet na Aurick hadt kunnen gaan.* 597.
——— *klaghten van den Paghter, te examineeren.* 636. 649.
——— *klaghten over wanbetalinge van den Norderpaghter, te examineeren.* 674.
——— *gelaſt ſijn klaghten aan het Furſtelycke Miniſterie op te geven, om ordres, te examineeren.* 723.
——— *rapport wegens difficulteyten in het invorderen der paghtpenningen en reſolutie.* 751.
Robiechtſen *Octroy om te diſponeeren.* 253.
la Rocque *notificeerende het voorgevallene op den laaſten Beededagh.* 240.
——— *gelaſt dien aangaande te procedeeren tegens de Contraventeurs.* 258.
——— *wegens patent voor ſijn Compagnie, de Raadt van Staate te adviſeeren.* 510.
Roelandts *wegens Utrecht gecommitteert in des Generaliteyts Reekenkamer.* 327.
Roeters *en Halmaal Brieven van voorſchryvens aan den Conſul Beeltſnyder.* 74.
——— *item aan den Heer vander Meer.* 475.
Regh, cum ſuis, *om uytwiſſelingh van Berkboat, de Admiraliteyt van Amſterdam te adviſeeren.* 79.
Romborgh *verſoeck om paſport, afgeweeſen.* 388.
Roſijn, *Weduwe, vyftigh guldens toegeleght.* 322.
de Roy *aangeſteldt als Secretaris van Nulandt.* 661.
van Roye *om pardon de Vice Hooghſchout Ghyſen te berichten.* 455.
de Roy, *Weduwe, ſes guldens toegeleght.* 8.
——— *verſoeck om alle drie maanden yets tot ſubſiſtentie, afgeweeſen.* 361.
——— *om aſſiſtentie, afgeweeſen.* 348.
de Rue *twee maanden verlof.* 632.
Rumpf, *Envoyé, advertentie.* 25. 37. 57. 79.

Figure 1. A column from a resolution page (left side) with meeting date, attendants list and resolutions, and a column from an index page (right side) with indexed terms and page references.

As the resolutions have a fixed format, this allows for segmenting the printed pages using a lot of fixed formulas indicating 1) the start of a new session, 2) a list of representatives attending the session, 3) a number of propositions, that end with an indication of the state of decision (accepted, rejected, or on hold). Finding these parts of text entailed using the paragraphs of the OCRed text and their visual characteristics to identify them. These visual clues were meant to guide contemporary users through the text and find important information quickly. Two examples illustrate this:

1. For the attendance list the capitalized words "PRAESIDE" for the president and "PRAESENTIBUS" for the common members. While these words often did not come out correctly in the OCR process, their place and capitalization made it easier to find them using approximate heuristic tools and in this way identify the attendance lists.
2. Similarly, to determine the chronologically ordered sessions across the pages, we used an iterative approach in which first a rough list of dates for sessions was identified (Table 1), which in a second pass could be completed for the missing sessions, that we then knew to be within a specific range of pages.

We treat the extraction of information as a text collation problem (Gilbert 1973): the resolutions have textual overlap when it comes to the start of a meeting day, the introduction of each individual resolution and its decision paragraph, with some unknown amount of textual variation (including variation in spelling, phrasing and text recognition errors), and the aim is to identify and align the textual repetitions and variations.

The formulaic expressions in the resolutions makes them good candidates for fuzzy searching. In this paper we discuss the methods we developed and the evaluation results of applying them on 100,000 pages and roughly 50 million words of resolution text of the printed resolutions from 1703 until 1796. This same principle was used on the entities in the resolutions. The important persons, institutions and subjects figuring in the resolutions were summarized in the index in the front of the book, with references to the pages. They give both the terms and the location where to find them in the text, which makes it much easier to find them using approximate matching.

Date string	Week day	Date	Page number	Paragraph
Lune den +. Januari 1725.	Lunae	1725-01-01	1	3
Martis den 2. Jannarii 1725.	Martis	1725-01-02	1	4
Mercuri: den 3. Januaris 1725.	Mercurii	1725-01-03	7	2
Jovis den 4. Jannarii 1725.	Jovis	1725-01-04	11	1
Sabbathi den 6. Januarii 1725.	Sabbathi	1725-01-06	14	7
Dominica den 7. Januari 1725.	Dominica	1725-01-07	19	1
Lana den 3. Januarij 1725.	Lunae	1725-01-03	19	1
Martis den 9. Janaarii 1725.	Martis	1725-01-09	21	4
Fovis den 11, Jaanarii! 1724.	Jovis	1725-01-11	28	1
Dominica den 14. Januarit' 1725.	Jovis	1725-01-14	33	9
Veneris den 19. Januarii 1725.	Veneris	1725-01-19	46	1
Lune den 22. Januarii 1725. vu	Lunae	1725-01-22	54	3
Jovis den 25. Jannarii 1725.	Jovis	1725-01-25	67	1
Veneris den 26. Januarii 1725.	Veneris	1725-01-26	70	3
Luna den 29 Januarii 1725.	Lunae	1725-01-29	76	6

Figure 2. Extraction of meeting date metadata and the page and paragraph number where the resolutions of that day start. Column 1 contains the meeting date as found in the OCR output.

Adding Domain-Specific Semantics to Extracted Information

The fuzzy searching algorithm takes as input a manually created phrase model, which is a list of keywords and phrases of interest, where each entry can have an optional list of alternative phrases or spellings, and uses character skip-grams to find candidate strings in the text for each phrase in the model. The searcher can be configured with different thresholds for edit distance and length variations (candidates may be shorter or longer than the phrase in the model).

We used fuzzy string searching to identify formulaic expressions and iteratively built a corpus-specific phrase model with which we identify:

1. the date and attendance list of each meeting, which are followed by all the resolutions of that day,
2. resolution boundaries, e.g. where they start and stop in the running text, so we know which text belongs to which resolution,
3. different types of opening phrases that correspond to different types of sources (e.g. requests, missives, reports, etc., see Figure 2), and
4. the decision paragraphs that state what decision, if any, was reached.

By recognizing the opening of a meeting, we know that the mentioned date is the meeting date, which allows us to not only recognize the date, but also adding a label for what the date refers to.

The attendants lists follow the opening of the meeting, and are structured. The first person mentione d is the president, so the recognized name can be labelled with the role of president. The function of president rotated by week among the provinces and presidents therefore also appear as common delegates in other weeks, where they can be found. This leaves a number of unidentified names to spot. All attendants are grouped by province, so can be labelled with their role as attendant of the meeting and with the province they represent. We have lists of all delegates in the meetings of the States General including years and province, that can be used to identify delegates. As a bonus, many delegates also appear in the resolutions in other roles as commissioners or representatives of the States.

year-1725-scan-507-even-para-3
Mercarii den 5. December 1725. PRESIDE, Den Heere Vegilin. PRASENTIEUS, De Heeren Van Lynden van Welderen, van Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. Van Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris Gedeputeerde uyt de Provincie van

year-1725-scan-507-even-para-3
Mercarii den 5. December 1725. PRESIDE, Den Heere Vegilin. PRASENTIEUS, De Heeren Van Lynden van Welderen, van Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. Van Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris Gedeputeerde uyt de Provincie van

year-1725-scan-507-even-para-3
Mercarii den 5. December 1725. PRESIDE, Den Heere Vegilin. Vegilin. PRASENTIEUS, De Heeren Van Lynden v an Welderen, v an Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. an Gelderlan dt. Van Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris

year-1725-scan-507-even-para-3
Mercarii den 5. Decem ber 1725. PRESIDE, Den Heere Vegilin. Vegilin. PRASENTIEUS, De Heeren Van Lynden an Welderen, v an Dam, Umbgroeven met cen extraordinaris Gedeputeerde uyt de Provincie van Gelderlandt. an Gelderlan dt. Van Maasdam Maasdam Steyn, van Marfeveen, Boon, met een extraordinaris Gedeputeerde uyt de Provincie van Hollandt en WeftVrieslandt. Welters, elters, Ockerfe , Noey , van Hoorn. Taats van Amerongen an Renswoude , met een extraordinaris Gedeputeerde uyt de Provincie van Utrecht.

Figure 3. Subsequent stages of spotting delegates in the attendance lists

We can add these elements as meaningful metadata labels to individual resolutions, which aids subsequent information access and analysis. Moreover, we can use the consistency of the structure of the resolutions to spot errors. For instance, we can check for cases where an opening formula was found but no decision formula. Or check for missing meeting dates by temporally ordering the ones we did find, or check for OCR errors in delegate names that occur often. In other words, the 'messy' output of the OCR process is turned into what Christoph Schöch (2013) calls 'smart data', i.e. cleaned, structured and semantically explicit layers of metadata.

Results and Evaluation

We built ground truth datasets to evaluate the different elements of our extraction pipeline: 1) traditional NER using only the resolution text to demonstrate the problems encountered with conventional NLP tools, 2) the identification of page types (index page, resolution page, title page), and 3) the phrase model and the fuzzy searching and extraction process for identifying the meeting dates and the individual resolutions and decisions.

For NER, we labelled all named entities in 200 manually transcribed resolution pages. We re-trained the Spacy NER tagger for Dutch using 10-fold cross validation, taking 90% of the pages for training in each fold, and the remaining 10% for testing. Even though these pages have no OCR errors (but no doubt some transcription errors), the best training run resulted in an F1 score no higher than 0.28 (with a precision of 0.49 and recall of 0.19). Although this can probably be improved with more training data, we conclude that this traditional NLP approach is not suitable for providing high-quality information access.

For the page type identification, we randomly sampled 1696 scans with 3376 pages (most scans contain two pages, some only a single front or back cover) and manually annotated the page type of each page. In the first step, we used layout analysis and fuzzy matching of our phrase models, which achieved an accuracy of 0.91. In the second step, we used the title pages to identify different parts of each book (e.g. the section of index pages, the section of resolution pages) and in each section chose the most frequent page type as the correct page type for all pages in that section. This increased the accuracy to 0.99.

For the ground truth for meeting dates and resolutions, we randomly sampled 300 historical dates in the period 1703-1796, locating the pages containing the resolutions for those dates, and manually annotated

the opening of the meeting, the attendance list and all resolution openings and decisions. For the meeting dates and attendance lists, our domain model and fuzzy searching approach reached a precision of 99% and recall of 93%. For the resolution openings and decisions, we reached a precision of 90% and recall of 68%. We have not yet modelled insertions of extracts and letters, and therefore have not evaluated those aspects yet. We are in the process of building a ground truth dataset for the attendants lists.

Reusability and Generalisability

Extracting semantic information requires heuristics for specific digitized resources. Generic approaches of layout analysis can detect standard structures like tables, figures, footnotes, headers and tables of content (Doermann and Tombre 2014, Clausner 2019), but cannot interpret specific semantics such as temporal orderings of meeting dates and the geographical ordering in the attendants lists.

The page identification rules and phrase model we developed in the Republic project are specific for the Resolutions of the States General, but the general method is reusable. Many large-scale resources have an internal structure that was created to enhance accessibility. Often, they were part of an information system designed for an analog, or 'paper' age, but it pays off to devise ways to transfer this to a digital equivalent. Devices like indexes contain a wealth of information about a resource, its context and its use. In many official resources corpus-specific phrases are used that may be categorized and used for searching. As there is often a degree of variation, the extraction process requires fuzzy searching and matching. Our fuzzy searching module is configurable, with matching thresholds that can be adapted to the peculiarities of the corpus or even to specific parts of the corpus. Fuzzy searching can also help to overcome the limitations of inaccurate OCR and HTR.

Beyond the Republic project, we have used this approach for a number of projects. For example, we have used it for books with medieval charters, in which the standard charter structure allows us to date the 17,000 historical place names mentioned in those charters and use them as place name attestations in historical gazetteers. Another example is the standard structure of advertisements of auctions in 18th century Dutch newspapers, in which brokers, auction date and venue and the auctioned goods can be extracted with a simple phrase model despite high character error rates.

A limitation of our method is that it does not work out-of-the-box but requires a corpus-specific phrase model, as existing information systems and structures are specific to the resource and require information extraction techniques to be adapted to corpus. This requires an iterative approach in which the models will be evaluated and refined. Standard phrases occur in many corpora, but they are usually corpus-specific. Furthermore, even this approach requires a minimum quality of text recognition to have enough contextual information for recognizing phrases.

Our method is still in development, but we think it is possible to make a toolkit with components that are relevant and reusable for many corpora. For different corpora, using the toolkit will be a matter of tuning the components. However, this must be done programmatically; we think that using an approach using just graphical user interfaces would obscure too many of the features and the implications of tuning. For evaluation of intermediate results in the iterative process, however, existing tools can be used.

Understanding Depth in Deep Learning: Knowledgeable, Layered, Impenetrable

Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Lauren Tilton
University of Richmond, United States of America

Keywords: critical data studies; algorithmic bias; neural networks; computer vision
Topics: Global, English, Contemporary, artificial intelligence and machine learning, image processing and analysis, Computer science, Media studies

We begin our paper by interrogating of the concept of "depth" within deep learning. We argue that the term has a triple meaning: *knowledgeable*, the accuracy displayed in the model's ability to excel in certain machine learning tasks, *layered*, a visualization of the learned hierarchical structures (Figures 1-2), and *impenetrable*, the inherent lack of interpretability and understanding (such as in the "deep sea" or "deep space") of their algorithmic operations. By illustrating the implications of each of these meanings, we show that all three are intricately linked to each other. Building off of the unavoidable interdependence between these elements of deep learning, several concrete outcomes

emerge from our characterization of deep learning models: (1) The need to introduce the concept of a deep problem, (2) treating embeddings as objects of study, and (3) the need to train scholars from a wide range of fields in the technology of deep learning.

Opening the Stage -- A Quantitative Look at Stage Directions in German Drama

Peer Trilcke
trilcke@uni-potsdam.de
University of Potsdam

Christopher Kittel
Open Knowledge Maps

Nils Reiter
nils.reiter@ims.uni-stuttgart.de
University of Cologne University of Stuttgart

Daria Maximova
National Research University Higher School of Economics, Moscow

Frank Fischer
National Research University Higher School of Economics, Moscow

Keywords: drama, stage directions, history of literature, corpus exploration
Topics: Europe, English, 18th Century, 19th Century, 20th Century, natural language processing, text encoding and markup language creation, deployment, and analysis, Literary studies

1. Introduction

Stage Directions (SD) have been stated as a "particularly underdeveloped topic" (Aston, Savona 1991: 71), even if this seems to be changing (cf. Aebischer 2003; Detken 2009; Dustagheer, Woods 2017; Habscheid et al. 2018). However, there are hardly any quantitative studies of SD so far (except of Sperantov 1998 who analysed Russian tragedies).

Usually, dramas consist of two types of text: Spoken Text (ST) and SD, where SD often consist of rather formal instructions. To complement existing quantitative approaches to drama, we will profile SD as an object of research in digital literary studies. To this purpose, we will give an exploratory analysis of quantitative features of SD using 384 German plays. Subsequently, we will characterise the development of SD over a period of 200 years, using the hypothesis of a "tendency towards epification" ("Episierungstendenz", Szondi 1956) as an interpretative framework.

2. Corpus and Workflow

Our corpus is based on the drama collection GerDraCor (Fischer et al. 2019). Even though literary studies has not yet reached a consensus on a definition of SD (Dustagheer, Woods 2017: 1-16), there are relatively uncontroversial (typographical) conventions what should be regarded as SD (Tonger-Erk 2018: 431-434). Digital editions using TEI (including GerDraCor) encode SD using the element.

Of the 474 plays available in GerDraCor, we removed librettos and 3 plays without SD, which yields a corpus of 384 plays that are pre-processed using the DramaNLP package[1]. The package employs the Stanford PoS tagger (Toutanova et al. 2003; STTS tagset: Schiller et al. 1999) and mate lemmatizer (Björkelund et al., 2000) separately on the text types ST and SD. The resulting data is analysed with the DramaAnalysis R package (Reiter 2019).

3. Exploratory Data Analysis

The median number of tokens is 1.639 in SD and 20.270 in ST (Fig. 01), so on average SD takes up 7.48% of the entire text. Additionally, sentences in SD are on average only about half as long as in ST.

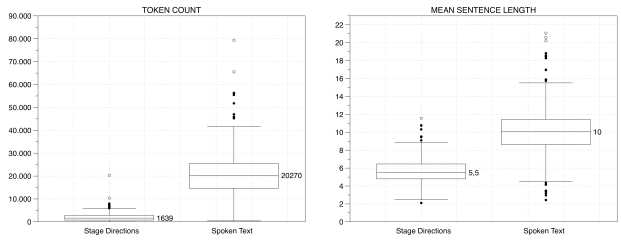


Figure 1: Distribution of token count and mean sentence length

SD and ST also differ in the proportions of certain parts of speech (POS). While SD has a higher percentage of adjectives and named entities, ST has a higher percentage of verbs and pronouns (Fig. 02).

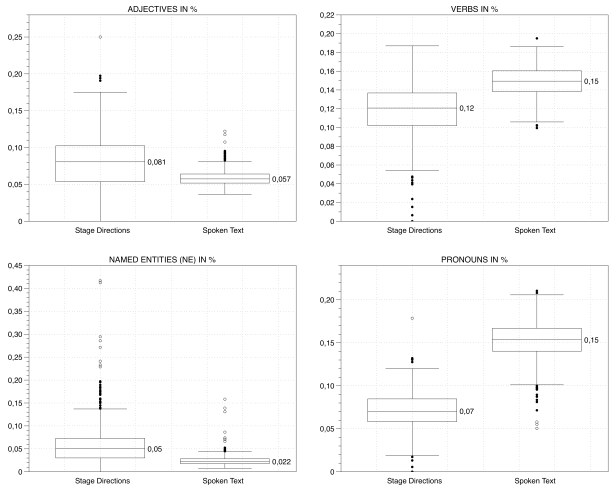


Figure 2: Portion of selected POS[2]

A closer look can be taken by differentiating fine-grained POS tags of adjectives and verbs (Fig. 03). Compared to ST, SD are about named entities (NE) doing something (VVFIN) in a specific way (ADJD), while they lack the rhetorics of addressing (PRO) and modality (VMFIN), which are more characteristic for ST.

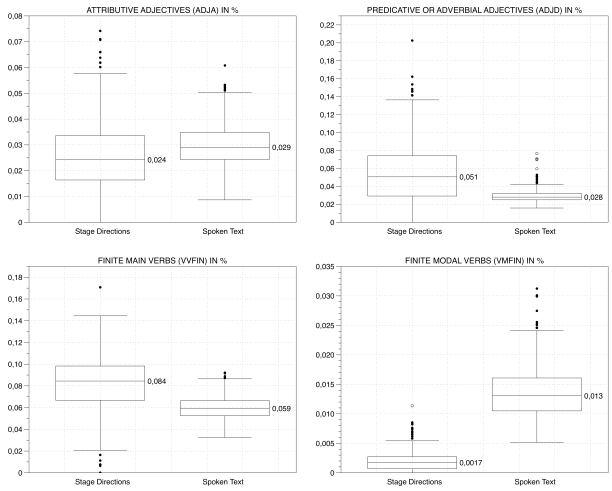


Fig. 03: Portion of selected fine-grained POS

4. Epification

While an intense discussion about (Brechtian) "Epic Theater" emerges in the aftermath of the 'crisis of drama' around 1890, "epification" in recent narratological and semiotical studies is conceptualised as a specific "mediation of the representation of events by a narrator-like instance" (cf. Weber 2017: 209). In contrast to the SD of 'classical' drama, primarily functioning as an "instruction" (Issacaroff 1988), the epified SD "seems to be incorporating qualities of the novel and the plastic arts" (Suchy 1991:80).

Consequently, we characterise the development of SD in our corpus with respect to such an understanding of epification, with five concrete indicators:

- the proportion of SD in the total text increases, giving more 'word space' to a potential narrative voice;
- sentences in SD get longer, thus becoming similar to SP;
- SD and ST converge in their POS distributions, so that SD approaches ST in the 'verbal gesture';
- words describing a location increase in SD, indicating a more concrete shaping of the diegetic world by a narrative voice;
- SD becomes more emotional, indicating an increase in evaluations and (possible) internal focalisations, both understood as privileges of the narrator's voice.

While there is almost no change in the number of tokens in ST, their number in SD increases over time, leading to a substantial increase in the SD/ST ratio (Fig. 04). The same holds for the mean sentence length in SD, which in several dramas 'around 1900' even considerably exceeds the mean sentence length in ST (i.e. ratio over one; Fig. 05).

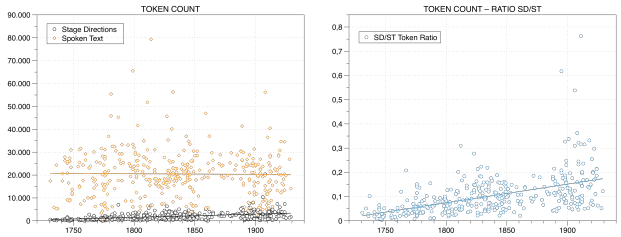


Figure 04: Historical distribution of token count

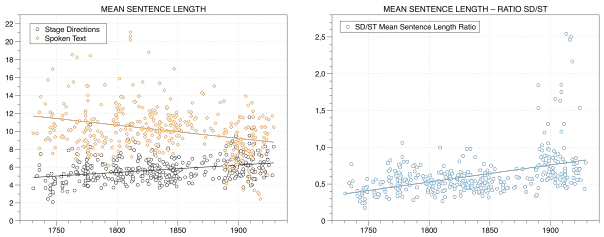


Figure 05: Historical distribution of mean sentence length

While verbs (Fig. 02) and attributive adjectives (Fig. 03) are synchronously more frequently found in ST, historically differentiated data shows a convergence of the percentages in SD and ST (Fig. 06). Conversely, the proportion of named entities in SD decreases (Fig. 06). Also noteworthy is the convergence of the proportion of relative pronouns in SD and ST (Fig. 06), which suggests

an increase in sentence complexity in SD. Overall, SD and ST become more similar over time.

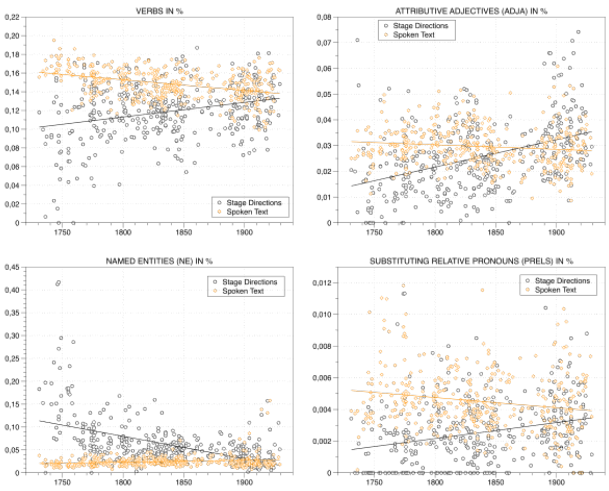


Figure 06: Historical distribution for selected POS

In order to explore 'topologisation', we defined a lexicon of topological terms. 3 As Fig. 07 shows, a considerable increase in topological words in SD can be observed, which can be interpreted as a more concrete topological order of the diegesis on the stage.

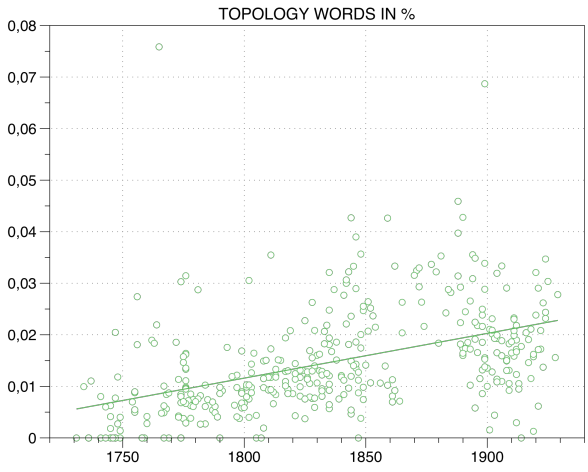


Figure 07: Historical distribution for topological words

Finally, using word-related information from GermaNet 11 (Hamp/Feldweg 1997), we calculated the percentage of words (adjectives, nouns, verbs) from the semantic field 'GEFUEHL' (emotions). As Fig. 08 shows, there is also a clear increase in such words in SD.

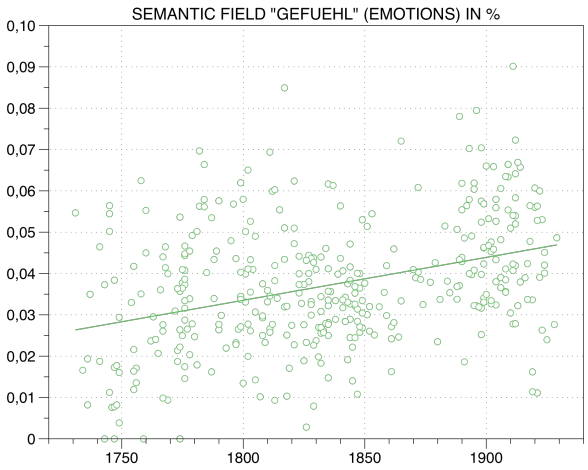


Figure 08: Historical distribution for 'emotion' words

5. Conclusion

Our examinations outlined SD as a relevant object of research for a quantitative history of drama. We demonstrated that SD is a text layer that differs from ST in several aspects. By examining the historical development of SD, we have shown a transformation of SD between 1730 and 1930, which we linked back to the "tendency towards epification" in German drama. Further research should develop more complex linguistic operationalisations of epification, and be extended to plays in other languages.

Bibliography

Pascale Aebischer (2003). Didascalia and Speech in the Dramatic Text. In: Journal of Dramatic Theory and Criticism 17.2 (2003), pp. 25-44.

Elaine Aston, George Savona (1991). Theatre as Sign System. A Semiotics of Text and Performance. London, New York: Routledge.

Anders Björkelund, Bernd Bohnet, Love Hafdell,Pierre Nugues (2010). A high-performance syntactic and semantic dependency parser. In: Coling 2010: Demonstrations, pp. 33-36, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL: http://www.aclweb.org/anthology/C10-3009.

Anke Detken (2009). Im Nebenraum des Textes Regiebemerkungen in Dramen des 18. Jahrhunderts. Tübingen: Niemeyer 2009.

Sarah Dustagheer, Gillian Woods (eds.) (2017). Stage Directions and Shakespearean Theatre. Bloomsbury: London, New York.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, Peer Trilcke (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In:

Proceedings of DH2019: "Complexities". Utrecht. URL: https://dev.clariah.nl/files/dh2019/boa/0268.html.

Stephan Habscheid, Constanze Spiess, Hartmut Bleumer, Niels Werber (eds.) (2018). Hauptsache Nebentext. Regiebemerkungen im Drama. LiLi - Zeitschrift für Literaturwissenschaft und Linguistik. Vol. 48, no. 3 (September 2018). Metzler, Stuttgart 2018.

Birgit Hamp, Helmut Feldweg (1997). GermaNet-a Lexical-Semantic Net for German. In: Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications on the conference of the Association of Computational Linguistics (ACL), 1997.

Michael Issacaroff (1988). Stage Codes. In: Performing Texts, ed. Michael Issacharoff, Robin F. Jones. U of Pennsylvania P, Philadelphia, pp. 59-74.

Nils Reiter (2019). DramaAnalysis 3.0.0. Available in CRAN. DOI: 10.5281/zenodo.3250613.

Anne Schiller, Simone Teufel, Christine Stöckert, Christine Thielen (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Annotation guidelines. URL: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf.

V. V. Sperantov (1998). Poetika remarki v russkoy tragedii XVIII - nachala XIX vv. (K tipologii literaturnykh napravleniy) [The Poetics of Stage Directions in Russian Tragedy of the Eighteenth and Early Nineteenth Centuries (On the Typology of Literary Movements)]. In: Philologica, vol. 5, no. 11/13 (1998), pp. 9-48. URL: https://rvb.ru/philologica/05/05sperantov.htm.

Patricia A. Suchy (1991). When Words Collide: The Stage Direction as Utterance. In: Journal of Dramatic Theory and Criticism 6.1 (1991), pp. 69-82.

Peter Szondi (1956). Theorie des modernen Dramas. Suhrkamp, Frankfurt.

Lily Tonger-Erk (2018). Das Drama als intermedialer Text. Eine systematische Skizze zur Funktion des Nebentextes. In LiLi - Zeitschrift für Literaturwissenschaft und Linguistik. Vol. 48, no. 3, pp. 421-444.

Kristina Toutanova, Dan Klein, Christopher Manning, Yoram Singer (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL, pp. 252-259.

Alexander Weber (2017). Episierung im Drama. Ein Beitrag zur transgenerischen Narratologie. De Gruyter, Berlin/Boston.

Notes

[1] https://github.com/quadrama/DramaNLP; doi: 10.5281/zenodo.3451307

[2] "ADJECTIVE" = ADJA, ADJD; "VERB" = VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP, VVFIN, VVIMP, VVINF, VVIZU, VVPP; "PRONOUNS" = PDAT, PDS, PIAT, PIDAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF.

StarCoder A general neural ensemble technique to support traditional scholarship, illustrated with a study of the post-Atlantic slave trade

Thomas Lippincott
tom@cs.jhu.edu
Johns Hopkins University, United States of America

Keywords: machine learning, history, linked data, collaboration
Topics: Global, English, North America, 18th Century, 19th Century, artificial intelligence and machine learning, data modeling, Computer science, History

We introduce StarCoder, a graph-aware neural autoencoder modell that can be easily used by traditional scholars to leverage current state-of-the-art machine learning architectures. StarCoder has the specific goals of working on a broad range of data without modification, while allowing computer science researchers to easily extend and specialize its behavior. We describe StarCoder and its early successes in support of an ongoing study of the post-Atlantic slave trade.

When Classical Chinese Meets Machine Learning: Explaining the Relative Performances of Word and Sentence Segmentation Tasks

Chao-Lin Liu
chaolin@g.nccu.edu.tw

National Chengchi University, Taiwan

Chang-Ting Chu
National Chiao Tung University, Taiwan

Wei-Ting Chang
University of California, Riverside, CA, USA

Ti-Yong Zheng
National Chengchi University, Taiwan

Keywords: explainable artificial intelligence, word segmentation, sentence segmentation, classical Chinese, machine learning
Topics: Asia, English, 5th-14th Century, Contemporary, artificial intelligence and machine learning, natural language processing, Asian studies, Humanities computing

When Classical Chinese Meets Machine Learning: Explaining the Relative Performances of Word and Sentence Segmentation Tasks

Chao-Lin Liu[1], Chang-Ting Chu, Wei-Ting Chang, Ti-Yong Zheng

Understanding classical Chinese is crucial for Chinese studies, partially because, in terms of time span, the majority of historical records about China were written in classical Chinese. Creating digital texts of the original documents either by optical record recognition or by human typists is required for employing digital assistance in Digital Humanities. The original Chinese texts usually did not include delimiters between words and sentences. Figure 1 shows a tomb biography that Liu and Chang used for discussing sentence segmentation of classical Chinese.[2] Unlike English, Chinese words are not separated by spaces, and there are no punctuation marks in the source texts. Word segmentation and sentence segmentation are thus important bedrocks for algorithmic understanding of classical Chinese, and the literature in digital humanities has started to see relevant discussions. Researchers applied statistical,[3],[4] machine-learning,[5] or deep-learning[6] methods to train computational models for the segmentation tasks.

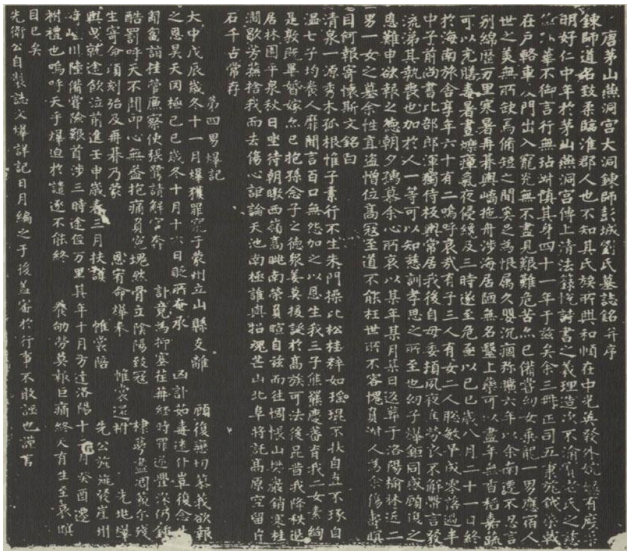


Figure 1: A sample tomb biography in classical Chinese

In standard machine-learning approaches, researchers acquire a collection of corpora for training and test purposes. Over the past many years, many, though not all, documents and books that were written in classical Chinese have been manually typed and saved as computer files. Modern punctuation marks have been added to the files at the same time. It is thus relatively easier to use the annotated texts for sentence segmentation than for word segmentation because it is very rare, if any, for typists to mark word boundaries in digitized texts. The training sets in Wang et al.'s, Han et al.'s and Shi et al.'s works have 237, 44, and 50 million characters, respectively.[7],[8],[9]

Although we expect that using more training data should lead to better test results in general, this beneficial effect is not automatic. Due to the nature of machine-learning experiments, especially of deep-learning ones, it is not easy (or appropriate) to compare experimental results directly by their numbers. Various factors may influence the quality of segmentation. The best F1 measure observed in Wang et al.'s work is slightly better than 80%.[7] When using a much larger training dataset for Qing, Han et al. did not report better results than when they used a much smaller dataset for the Tang epitaph.[8] Shi et al. defined their own measure for quality, and observed that the quality of segmentation varied between mid 60% and mid 90% in experiments for data of different Chinese dynasties.[9]

Machine-learning based methods provide us effective ways to build a running system for sentence segmentation. When provided with sufficient data, we may train computational models for sentence segmentation, and achieve impressive results.[2],[7],[8],[9] Despite this convenience, previous research seldom

aims at investigating the factors behind the success and perhaps failure of applications of deep-learning techniques for segmenting classical Chinese . Some pioneers had discussed the needs of explainable artificial intelligence as early as 2004,[10],[11] and the demands are strengthening in recent years.[12] The quest for explanation not only satisfies our curiosity but also helps us build better systems for the segmentation tasks.

We may design different procedures to train computation models for sentence segmentation and observe the differences in the resulting quality of segmentation. The differences in the training procedures may provide clues as to what the computation models might learn from the training data. Though experimental findings alone are not solid proofs, the observations can be enlightening.

The table in Figure 2 provides some basic statistics about the corpora that were used in our experiments.[13] These corpora belonged to two genres and are representative documents of the Tang dynasty.[14] MZM contains only tomb biographies, and OTB and NTB are historical documents. To make the experimental results comparable, we split each of these corpora, and used 70% and 30% of each for training and test purposes, respectively.

Corpus	Number of Characters (NOC)	Number of Punctuations (NOP)	Ratio (NOP/NOC)
MZM (Tang Tomb Biographies, 唐代墓誌彙編)	2421077	446937	18.46%
OTB (Old Tang Book, 舊唐書)	1923361	359955	18.71%
NTB (New Tang Book, 新唐書)	1559655	321131	20.59%

Figure 2: Basic statistics of three corpora of classical Chinese (Tang dynasty)

We adopt a typical deep learning procedure for marking sentence boundaries in these corpora.[15] The corpora took turns to serve as training and test data. We produced the character embedding vectors of the corpora using Gensim,[16] and trained deep neural networks that are comprised of five layers of BiLSTM units in Keras.[17] We measured the quality of sentence segmentation with the F1 measure.

The table in Figure 3 shows the observed F1 measures for part of a system of experiments.[18] In the first experiment (ID=1), we used only MZM for training and test, and achieved 0.8772 in F1. In Experiment 31, we used MZM, OTB, and NTB for training the LSTMs to segment the test portion of MZM. With the help of the extra training data, the F1 increased slightly.

ID	Word2Vec			Training LSTM Stack			Test	F_1	ID	Word2Vec			Training LSTM Stack			Test	F_1
	M	O	N	M	O	N				M	O	N	M	O	N		
1	M			M			MZM	**0.8776**	40	M	O	N		O	N	MZM	0.5858
2		O			O		OTB	**0.7883**	41	M	O	N		O	N	OTB	**0.8545**
3			N			N	NTB	0.7907	42	M	O	N		O	N	NTB	0.8246
31	M	O	N	M	O	N	MZM	**0.8824**	43	M	O	N	M			MZM	0.8769
32	M	O	N	M	O	N	OTB	0.8659	44	M	O	N	M			OTB	**0.5695**
33	M	O	N	M	O	N	NTB	0.8325	45	M	O	N	M			NTB	0.4964
34	M	O	N	M	O		MZM	0.8822	46	M	O	N		O		MZM	0.5909
35	M	O	N	M	O		OTB	**0.8467**	47	M	O	N		O		OTB	**0.8325**
36	M	O	N	M	O		NTB	0.7480	48	M	O	N		O		NTB	0.7413
37	M	O	N	M		N	MZM	0.8759	49	M	O	N			N	MZM	0.4881
38	M	O	N	M		N	OTB	0.7778	50	M	O	N			N	OTB	**0.7535**
39	M	O	N	M		N	NTB	0.8047	51	M	O	N			N	NTB	0.7921

Figure 3: Using different combinations of corpora to generate word vectors and train classifiers

In Experiment 47, we produced character embedding with MZM, OTB, and NTB, trained the LSTMs with OTB to segment OTB, and achieved a better F1 than that for Experiment 2. In Experiment 41, we added NTB to train the LSTMs, and improved the F1. In contrast, in Experiment 35, we observed a smaller improvement when we added MZM to train the LSTMs. One possible reason for the reduction in improvement is that OTB is more similar with NTB than with MZM.

Another evidence for supporting this conjecture is that we achieved 0.7535 in Experiment 50 when we trained the LSTMs with NTB to segment OTB. When we used only MZM to train the LSTMs to segment OTB, the F1 was only 0.5695 in Experiment 44. When we do not train the LSTMs with OTB to segment OTB, we achieved much better results when training the LSTMs with a more related corpus.

We have extended our work on sentence and word segmentation that were reported in DH2018 and DH2019. On the computing side, we have achieved better results by switching from traditional machine-learning methods to deep-learning techniques. As a study in Digital Humanities, we attempt to find qualitative explanations for the changing improvements which were influenced by the varying combinations of training data. We analyzed our observations based on results observed in a large number of experiments statistically. We could provide just one sample discussion for sentence segmentation tasks here, and we wish to have an opportunity to discuss more details about our experience in DH2020.

Endnotes

[1] The authors are affiliated with the National Chengchi University in Taiwan, when they work on this research. Chao-Lin Liu is the corresponding author, and can be reached at chaolin@g.nccu.edu.tw. He is the president of the Taiwanese Association for Digital Humanities, a distinguished professor at the National Chengchi University, and the chairperson

of the Department of Computer Science. Chang-Ting Chu is a graduate student at the National Chiao Tung University in Hsinchu, Taiwan, and can be reached at ericzx103703012@gmail.com. Wei-Ting Chang is a graduate student at the University of California, Riverside, in USA, and can be reached at wz840427@gmail.com. Ti-Yong Zheng is a graduate student of the National Chengchi University, and can be reached at ccs96307@gmail.com.

[2] Chao-Lin Liu and Yi Chang (2018) Classical Chinese sentence segmentation for tomb biographies of Tang dynasty, in Proceedings of the 2018 International Conference on Digital Humanities, 231?235.

[3] Ke Deng, Peter K. Bol, Kate J. Li, and Jun S. Liu (2016) On the unsupervised analysis of domain-specific Chinese texts, in Proceedings of the National Academy of Sciences, 113(22):6154-6159.

[4] Chao-Lin Liu (2019) Onto word segmentation of the Complete Tang Poems, in Proceedings of the 2019 International Conference on Digital Humanities, 2019.

[5] Ethem Alpaydin (2014) Introduction to Machine Learning, third edition, Cambridge: The MIT Press.

[6] Ian Goodfellow and Yoshua Bengio and Aaron Courville (2016) Deep Learning, Cambridge: The MIT Press.

[7] Boli Wang, Xiaodong Shi, Zhixing Tan, Yidong Chen, and Weili Wang (2016) A sentence segmentation method for ancient Chinese texts based on NNLM. in Proceedings of the Chinese Lexical Semantics Workshop 2016, Lecture Notes in Computer Science 10085, 387-396.

[8] Xu Han, Hongsu Wang, Sanqian Zhang, Qunchao Fu, and Jun Liu (2019) Sentence segmentation for classical Chinese based on LSTM with radical embedding, The Journal of China Universities of Posts and Telecommunications, 26(2):1?8. (in Chinese)

[9] Xianji Shi, Kaiqi Fang, Xianjiong Shi, Xianju Shi, Xiandiao Shi, Xianji Shi, Xianda Shi, Xianfeng Shi, and Yanchun Song (2019) A method and implementation of automatic punctuation, Journal of Digital Archives and Digital Humanities, 3:1?19. (in Chinese)

[10] Michael van Lent, William Fisher, and Michael Mancuso (2004) An explainable artificial intelligence system for small-unit tactical behavior, in Proceedings of the Sixteenth Conference on Innovative Applications of Artificial Intelligence, 900-907.

[11] Mark G. Core, H. Chad Lane, Michael van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg (2006) Building explainable artificial intelligence systems, in Proceedings of Eighteenth Conference on Innovative Applications of Artificial Intelligence, vol. 2, 1766-1773.

[12] Daniel S. Weld and Gagan Bansal (2019) The challenge of crafting intelligible intelligence, Communications of the ACM, 62(6):70-79.

[13] Due to the word limit, we could not provide complete details about how we extracted the texts from the original documents to obtain these corpora. MZM is the complete contents of the two books of Tang Tomb Biographies (tang2 dai4 mu4 zhi4 hui4 bian1) and one book of sequel (tang2 dai4 mu4 zhi4 hui4 bian1 xu4 ji2). OTB contains the biographies (lie4 chuan4) in the Old Tang Book (jiu4 tang2 shu1), and NTB contains the biographies (lie4 chuan4) in the New Tang Book (xin1 tang2 shu2).

[14] Tang is a Chinese dynasty that existed between 618CE and 907CE.

[15] Note that the step for word embedding and the step for training the LSTM units are separate.

[16] Gensim is available at https:// radimrehurek.com /gensim/. When using word2vec to create the embedding vectors for Chinese characters, we set window (the window size) to 12, min_count to 1, batch_words to 8000, iter to 50, and produced vectors of 300 dimensions. We only report results of using the skip-gram algorithm.

[17] Keras is available at https://keras.io/. Inherited from our work that was reported in DH2019, we consider a context of six characters for determining whether a character is the last character of a sentence in training our BiLSTM stack. The outputs of the BiLSTM units have 400 dimensions, the optimizer is Adam, and we used softmax as the activation function for the final output layer. The output layer is a Dense layer, which used the categorical_crossentropy as the loss function.

[18] In the table in Figure 3, M, O, and N denote MZM, OTB, and NTB in the table in Figure 2, respectively. The columns M, O, and N indicate whether the corpora were used in creating word-embedding vectors and whether the corpora were used in training the LSTM stack. To facilitate the comparisons, we used only one corpus in the tests. We showed results for part of our experiments, so there are missing IDs in the table in Figure 3.

"Telling bigger stories": Formal ontological modelling of scholarly argumentation

Alexander Huber

alexander@hubers.org.uk
University of Oxford, United Kingdom

Keywords: knowledge modeling, semantic web, graph technology, RDF, semantic analysis
Topics: Europe, , English, 18th Century, linked (open) data, network analysis and graphs theory and application, Literary studies

"Telling bigger stories": Formal ontological modelling of scholarly argumentation

Introduction and problem statement

This paper presents completed work on formalizing scholarly argumentation in literary studies in the form of *knowledge models*. It argues that by adopting formal ontological modelling[1], not only can we make scholarly argumentation conducive to long-term preservation and ongoing participation in the scholarly discourse, but also make it the starting point for *genuinely collaborative research* endeavours. Moreover, by virtue of their ontological foundation, knowledge models ideally represent *reproducible* interpretative processes.

Knowledge modelling has been an ongoing field of research in the digital humanities for many years. Flagship projects such as *ResearchSpace*[2] have helped to popularize the concepts of *semantic integration* and to make the benefits of this approach tangible for the research community. While there has been acknowledgement of the importance of formalization of scholarly argumentation in both literary studies and digital humanities[3], there have been a number of practical and ideological barriers to its realization. The former often relate to the perceived steep learning curve associated with ontologically underpinned knowledge modelling, the latter to a perceived infringement of scholarly expressivity and freedom. This paper addresses both concerns by integrating an easy-to-use modelling tool with a literary corpus, highlighting opportunities for beneficial formalisms, clarity, and precision of expression wherever possible, so users are enabled to collaboratively create and collectively benefit from improved *knowledge representations*.

In the context of the research strand of the Eighteenth-Century Poetry Archive (ECPA) project[4], I have developed a Web-based knowledge modelling tool intended to enable literary scholars to model their argumentation about the poems in the database.[5] I will demonstrate that as a form of integrated knowledge representation, visualization, and preservation, the knowledge models created can both *express and integrate* different (including contradictory) argumentative strands.

The tool thus encourages intersection of otherwise dispersed and isolated scholarly arguments, the breaking up of intellectual silos, and helps towards dispelling the persistent myth of the lone scholar in the humanities. In the spirit of the conference theme, the paper touches on many aspects of the open data movement that underpin and are prerequisites for this work, such as *Open Access, Open Scholarship,* and *Open Standards* (such as LOD, shared ontologies, taxonomies, and vocabularies).

Approach and implementation

The ECPA knowledge modelling tool facilitates and encourages the creation of models that focus on different aspects of a poem, yet are all collectively modelling the object of study and all arriving at ever more sophisticated and comprehensive models. It is through formalization using shared conceptualizations, that the models created remain useful beyond the original context and purpose of their creation. They can thus be *shared, re-used, adapted (forked), enhanced, aggregated, and even developed collaboratively.*

The ECPA modelling tool is both a modelling and publishing platform. As such, it offers a number of *workflow management* options that support the entire modelling process, from inception through to publication. It has been designed in a way that is easy to grasp, learn, and perform. Its layout consists of a number of workflow buttons (e.g. fork, close, save, download, publish) along with the model itself in both its *graph representation*[6] and a *control panel*, which provides information about the model, a representation of the model's components (classes, instances, and statements) in text form, and an editing tool.

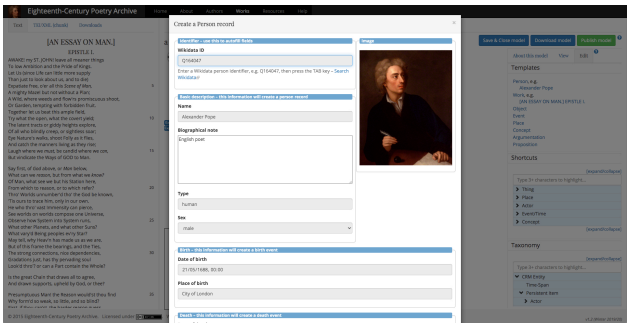


Figure 1: Template for a person record pre-filled with Wikidata information

The editing tool offers three related approaches to the creation of the instances (individuals) and relations that together form a scholarly argument. Firstly, a number of *templates* are provided for some of the key ontological classes frequently used in modelling (currently, persons, works, objects, events,

places, concepts, and argumentation). These templates can be auto-filled from Wikidata representations of the items where available. Secondly, the tool offers a number of *shortcuts*[7], which provide a convenient way of connecting common classes in standardized ways. Shortcuts are currently provided between things, places, actors, events/time, and concepts. Each shortcut takes the form of a path of one or more statements (triples) that follow on from each other. Thirdly, the full expressivity of the ontology is available by using a *taxonomy tree display* of the classes in the project's ontologies, which provides the most granular modelling option. The best results are often achieved by using the three approaches (*templates, shortcuts,* and *taxonomy*) in parallel.

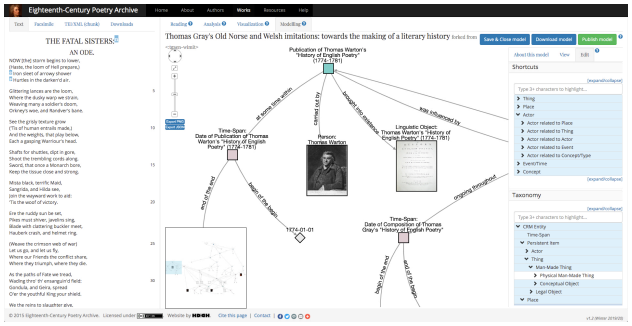


Figure 2: Part of a knowledge graph with ontology shortcuts and taxonomy tree in the editing view

The modelling process is augmented by a context-sensitive interface throughout, which offers definitions of classes, scope information, and help texts. As instances are added to the knowledge base, these are used to populate data lists that can be used to speed up the process of filling in the forms to create the triples.

Contribution and further work

The ECPA modelling tool presented here takes knowledge modelling as a method of semantic integration and applies it in an integrated manner to the literary study of a coherent corpus of eighteenth-century poetry, thus arriving at a fully-fledged implementation of a *research environment for literary scholars*[8], which enables the creation, representation, and long-term preservation of the scholarly analysis of the poems. Most importantly, as a participatory methodology to *produce and organize shared knowledge*, it facilitates new types of collaboration.

By making our research reproducible, extensible, and thanks to the ontological underpinnings, machine-understandable, we also embrace the machine's capacity for reasoning as an important way of producing new insights. Once a big enough set of models has been created, we can use computers to analyse them, reason upon them, and help us to not only store these knowledge bases, but build on them to create *new knowledge*. This step connects to ongoing efforts, primarily in scientific, legal, and medical contexts, to arrive at formalized argumentative models through *context-guided argumentation mining*. Underpinning these efforts with a common ontology should contribute to a set of coherently structured if unevenly granular knowledge representations. Ultimately, I expect these models to constitute a dataset that will be useful for the computationally-assisted *interpretation* of the poems. I hope to show that knowledge models are thus ideally suited not only to enhance our ability to ask bigger questions by semantically integrating different strands of scholarly arguments, but also through sharing and collaboration, aggregation and combination are capable of telling bigger stories.

The paper will demo the *research environment*, will cover the *technology layer* underpinning it (RDF store and library, triple generation, graph representation, and interface components such as serializations and storage formats), and will report on *outreach efforts* and *uptake*.

Notes

1. I have adopted the CIDOC-CRM family of ontologies as core ontologies for this work: http://www.cidoc-crm.org/collaborations
2. https:// www.researchspace.org/
3. As early as 1990 Helmut Bonheim has argued for the benefits of formalization in literary studies in his book *Literary Systematics* (Cambridge, 1990). The field of Digital Humanities has long acknowledged the centrality of knowledge modelling as one of its core activities, pertinent contributions by Willard McCarty, Øyvind Eide, and Fabio Ciotti, among others testify to its centrality.
4. Eighteenth Century Poetry Archive https:// www.eighteenthcenturypoetry.org/
5. There is no reason why this tool couldn't be published in a generic stand-alone fashion, however, its integration with a literary corpus as well as other tools such as close reading and visualization tools in ECPA provides benefits of contextualization not otherwise easily achievable.
6. Via Cytoscape.js http://js.cytoscape.org/
   These shortcuts are based on the CIDOC-CRM's idea of using "Fundamental Categories" (FC) and "Fundamental Relationships" (FR) for querying CRM- based repositories.
7. The ECPA project has been integrating digital tools into its textual corpus from inception, firmly believing that it is through integration of object and tool we achieve deeper contextualization and better outputs.

Bibliography

Bartalesi, Valentina and Carlo Meghini. "Using an Ontology for Representing the Knowledge on Literary Texts: the Dante Alighieri Case Study". Semantic Web 1 (2015): 1-10. Web. 29 Aug. 2018. http://www.semantic-web-journal.net/content/using-ontology-representing-knowledge-literary-texts-dante-alighieri-case-study-0

Beynon, Meurig, Steve Russ and Willard McCarty. "Human Computing- Modelling with Meaning". Digital Scholarship in the Humanities 21(2) (2006): 141-157. Web. 18 Aug. 2017. https://doi.org/10.1093/llc/fql015

Bonheim, Helmut. Literary Systematics. Cambridge: D. S. Brewer, 1990. Print.

Ciotti, Fabio. "Digital methods for Literary Criticism". Lecture slides. University of Rome Tor Vergata, 2015 Web. 29 Aug. 2018. http://didattica.uniroma2.it/files/scarica/insegnamento/161783-Informatica-Umanistica-Lm-Per-Il-Llea/37175-Slide

Ciula, Arianna and Cristina Marras. "Circling around texts and language: towards 'pragmatic modelling' in Digital Humanities". Digital Humanities Quarterly (DHQ) 10.3 (2016). Web. 18 Aug. 2017. http://www.digitalhumanities. org/dhq/ vol/10/3/000258/000258.html

Ciula, Arianna and Øyvind Eide. "Modelling in digital humanities: Signs in context". Digital Scholarship in the Humanities 32 (2017): i33-i46. Web. 18 Aug. 2017. https://doi.org/10.1093/llc/fqw045

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle, eds. Models and Modelling between Digital and Humanities - A Multidisciplinary Perspective. Historical Social Research (HSR) Supplement 31 (2018). Print.

Eide, Øyvind. Media Boundaries and Conceptual Modelling: Between Texts and Maps. Pre-print manuscript, Apr. 2015. Web. 7 Jan. 2019. https://www.oeide.no/ research /eideBetween.pdf

Kräutli, Florian and Matteo Valleriani. "CorpusTracer: A CIDOC database for tracing knowledge networks". Digital Scholarship in the Humanities 33(2) (2018): 336-346. Web. 26 Oct. 2018. https://pure.mpg.de/rest/items /item_2472866_10/ component/file_3002633/content

Lippi, Marco, and Paolo Torroni. "Argumentation Mining: State of the Art and Emerging Trends." ACM Transactions on Internet Technology 16, no. 2 (March 30, 2016): 1-25. https://doi.org/10.1145/2850417

McCarty, Willard. Humanities Computing. Houndmills: Palgrave Macmillan, 2005. 20-72. Print.

Oldman, Dominic, Martin Doerr, and Stefan Gradmann. "Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge". Schreibman, Susan, Ray Siemens, and John Unsworth, eds. A New Companion to Digital Humanities. Malden, MA: Wiley Blackwell, 2016. 251-273. Print.

Pasin, Michele and John Bradley. "Factoid-based prosopography and computer ontologies: towards an integrated approach". Digital Scholarship in the Humanities 30(1) (2015): 86-97. Web. 29 Aug. 2018. https://doi.org/10.1093/llc/fqt037

DH Awards: Possibilities and problems of an openly-nominated and openly-voted DH awareness-raising activity

James Cummings

james.cummings@newcastle.ac.uk
Newcastle University, United Kingdom

Keywords: DH Awards, DH Awareness, Community, Outreach, Voting
Topics: Global, English, Contemporary, digital activism and advocacy, digital ecologies and digital communities creation management and analysis, Education/ pedagogy, Humanities computing

This long paper explores the annual *Digital Humanities Awards*, run as an openly-nominated and openly-voted DH awareness-raising activity at http://dhawards.org/, through a history of their development and consideration of a number of problematic aspects concerning them in assessing their future. Although often praised, there has been a variety of criticism of them. Presented by the founder and main instigator of *DH Awards*, this paper opens up discussion on these issues, while benefiting from a privileged position of having access to all of the confidential raw nomination and voting data for the entire run of *DH Awards* (2012 to present). Additional feedback will also be collected during the *DH Awards 2019* round of voting (in early 2020). This paper does not seek to dismiss criticisms, but to engage with them to discuss whether the *DH Awards* are a positive force and the best ways to ameliorate any issues raised.

Distant, Deep, Thick Reading across Languages: Building towards Inclusive Digital Epistemologies of Text

Gabriele Salciute Civiliene
gabriele.salciute-civiliene@kcl.ac.uk
Department of Digital Humanities, King's College London, UK

Geoffroy Noel
geoffroy.noel@kcl.ac.uk
King's Digital Lab, King's College London, UK

Ginestra Ferraro
ginestra.ferraro@kcl.ac.uk
King's Digital Lab, King's College London, UK

Natalia Romanova
natasha.romanova@kcl.ac.uk
King's Digital Lab, King's College London, UK

Arianna Ciula
arianna.ciula@kcl.ac.uk
King's Digital Lab, King's College London, UK

Keywords: cross-linguistic computing, translation, distant reading, repetition strings
Topics: Comparative (2 or more geographical areas), Europe, English, 20th Century, Contemporary, data modeling, text mining and analysis, Translation studies, Central/Eastern European Studies

While distant reading in one language is relatively straightforward, computation across languages poses epistemo-semiotic challenges. The fundamental problem here is how to cut across many texts in different languages in a way that would yield rich data and narratives about translated texts as socio-cultural artefacts. What cross-linguistic computational text analysis has to overcome, among other things, are technological conformism, linguistic colonialism, and reductive interpretations akin to the lack of 'thick descriptions' observed in cultural anthropology (Geertz 1973). In this presentation, the team of the DRaL (Distant Reading across Languages) project will discuss the techniques of modelling translatorial choices with word frequencies and repetition strings as patterns of dynamic behaviour. The context-aware computing is a major concern that drives the project towards building a visually rich narrative as a form of distant/deep/thick reading which would be inclusive of minority languages and their role in the making of global literature.

Variorum Editions Across Text and Music: the Early Modern Songscapes Project

Raffaele Viglianti
rviglian@umd.edu
University of Maryland, United States of America

Scott Trudell
University of Maryland, United States of America

Sarah Williams
University of South Carolina, United States of America

Keywords: scholarly editing, data modeling, text encoding, music encoding, performance
Topics: Europe, English, 15th-17th Century, manuscripts description, representation, and analysis, scholarly editing and editions development, analysis, and methods, Literary studies, Musicology

Early Modern Songscapes (EMS) is an interdisciplinary project that aims to develop an intermedia online platform to support scholarly investigations early modern English song, by tracing individual lyrics and musical settings that moved through different textual and performance contexts in sixteenth- and seventeenth-century England. The first iteration of the project, now completed, has focused on Henry Lawes's 1653 Ayres and Dialogues, highlighting versions of the songs that have been published, co-located with image and sound materials relating to each individual song. The edition includes both verse and music notation, which are encoded using the Text Encoding Initiative and the Music Encoding Initiative

formats. A handful of ayres exist in multiple sources, in which cases, variants from each source are shown as part of the digital editions. This presentation will focus on the modeling, processing, and presentation of these variorum editions in particular, which we argue is applicable beyond this project.

PoetryLab. An Open Source Toolkit for the Analysis of Spanish Poetry Corpora

Javier de la Rosa
versae@linhd.uned.es
Universidad Nacional de Educación a Distancia (UNED), Spain

Álvaro Pérez
Universidad Nacional de Educación a Distancia (UNED), Spain

Salvador Ros
Universidad Nacional de Educación a Distancia (UNED), Spain

Elena González-Blanco
CoverWallet, Inc., Spain

Keywords: syllabification, scansion, enjambment, rhyme
Topics: Europe, English, Spanish, 18th Century, 19th Century, 20th Century, natural language processing, rhetorical analysis, Computer science, Linguistics

Introduction

The transmission of text in poetic form is a quasi universal aspect in the oral tradition of every culture [1]. The study of the poetic features of text, especially their rhythmic structure when forming verses, pertains to the different traditions, whose scholars established the rules that might govern poetry. Within this context, the POSTDATA Project [2] formalized a network of ontologies able to express any poetic expression and its analysis at the European level, enabling scholars all over Europe to interchange their data using Linked Open Data. However, varied research interests result in corpora that might not share the same facets of an analysis. To alleviate this concern and foster the completeness of the interchanged corpora, our team set out to build a software toolkit to assist in the analysis of poetry. This paper introduces PoetryLab, an extensible open source toolkit for syllabification, scansion (extraction of stress patterns), enjambment detection (syntactical units split in two lines), rhyme detection, and historical named entity recognition for Spanish poetry. Our toolkit achieves state of the art performance in the tasks for which reproducible alternatives exist.

Design Principles

Manuals for metrical analysis of Spanish poetry exist at least since the 18th Century [6], although the foundational work and subsequent refined guides for modern analysis would take another century to appear (see, e.g., [3, 4, 5]). Despite such a long and rich tradition, not many computational tools have been created to assist scholars in the annotation and analysis of Spanish poetry. With ever increasing corpora sizes and the popularization of distant reading techniques, the possibility of automating part of the analysis became very appealing. Although solutions exist, they are either incomplete (i.e., fixed-metre poetry, mostly hendecasyllables [7, 8, 9, 10]), not applicable to Spanish [7, 12], or not open nor reproducible [8]. These limitations guided the design of PoetryLab. At its core (see Figure 1), PoetryLab provides a compliant OpenAPI [15] that connects independent packages together. Built on top of the natural language processing framework spaCy [16], two Python packages perform scansion and enjambment detection, namely, Rantanplan [26] and JollyJumper.[1] In Spanish, some words are stressed depending on their function in the sentence, hence the need for a proper part of speech (PoS) tagger. AnCora [19], the corpus spaCy is trained on for PoS tagging of Spanish texts, splits most affixes thus causing some failures in the tags it assigns. To circumvent this limitation and to ensure clitics were handled properly, we integrated Freeling's affixes rules via a custom built pipeline for spaCy. The resulting package, spacy-affixes,[2] splits words with affixes before assigning PoS [20], and can be plugged in to a regular spaCy pipeline loading one of the statistical models for Spanish. This pipeline is the foundation for Rantanplan and JollyJumper, which are rule-based algorithms inspired by Ríos Mestre [25], Caparrós [5] and Navarro Tomás [24], and Quilis [22] and Spang [23], respectively.

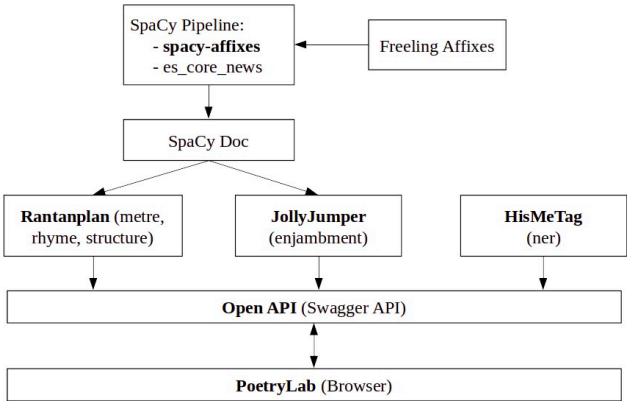


Figure 1. General architecture of PoetryLab.

Following the OpenAPI specification, we defined a REST API that unified the internal interface of the different packages and provided a common endpoint for analysis.[3] For external packages developed in languages other than Python, PoetryLab provides a pluggable architecture that allows their integration. This is the case for our named entity recognition system, HisMeTag [17], developed in Java and connected to the PoetryLab API through an internal REST API. The only requirement for third-party integrations is to consume text and produce both JSON and RDF triples.

The PoetryLab API was then used to provide with functionality a React-based web interface that non-technical scholars can use to interact with the packages in a graphical way (see Figure 2). The frontend also allows downloading the generated data.

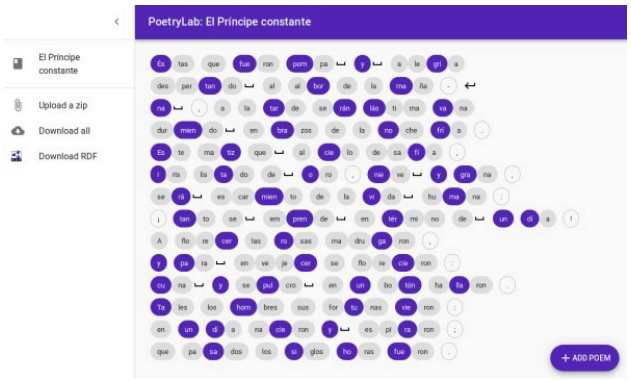


Figure 2. PoetryLab showing stressed syllables (blue), sinalefas (‿) and enjambments (↵).

Results

One notably difficult aspect of benchmarking automated analysis of Spanish poetry is the lack of a gold standard reference corpus. For the evaluation of the syllabification algorithm in PoetryLab we build a 100k words corpus using a combination of online resources[4], which we name EDFU and are releasing under a Creative Common license.[5] For metrical analysis we used Navarro-Colorado's corpus [21]. For mixed-metre we are using our own copus obtained from Carjaval's annotated anthology [11].[6] Unfortunately, we have not found a public corpus for rhyme and stanza identification yet, and although an enjambment corpus seems to exist [18], it is not publicly available.

Table 1 shows the ratio of success extracting the list of syllables of the words in EDFU, and the correct metrical analysis for the different corpora and tools. Notably, PoetryLab achieves state of the art performance for syllabification and per line metrical analysis.[7] We were unable to reproduce Gervás' approach and are reporting their own ratios.

Syllabification (EDFU)
- PoetryLab (rantanplan): 99.98
- Navarro-Colorado: 98.74
- Agirrezabal: 98.06

Metrical patterns (fixed-metre)
- PoetryLab (rantanplan): 96.22
- Navarro-Colorado:94.44
- Gervás:88.73
- Agirrezabal:90.84

Metrical pattern (mixed-metre)
- PoetryLab (rantanplan): 65.02
- Navarro-Colorado:49.38

Table 1. Summary of results of PoetryLab performance.

Discussion

Although at an early stage, PoetryLab has proven useful in that it highlights some issues with the existing corpora and techniques developed to this day. First, there was no alternative system to analyze poetry composed of other than hendecasyllables, for which we are using a corpus of mixed-metre poetry based on Carvajal's original annotations. Moreover, we are contributing with a new corpora to evaluate syllabification procedures, and enriching the ecosystem of Python tools for Spanish by providing a spaCy pipeline that deals with clitics. Finally, we make the data produced by the PoetryLab machine readable, interoperable, and ready to be ingested into a triple store compliant with the POSTDATA Project network of ontologies.

Eventually, PoetryLab will be integrated into the larger POSTDATA Project public website, making working with European repositories of poetry a more pleasant task, and assisting whenever possible with the metrical and rhetorical side of the analysis.

Founding Source

Research for this paper has been achieved thanks to the Starting Grant research project Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528) obtained by Elena González- Blanco.

This project is funded by the European Research Council (https://erc.europa.eu) (ERC) under the research and innovation program Horizon2020 of the European Union.

Notes

1. See https://github.com/linhd-postdata/rantanplan and https://github.com/linhd-postdata/jollyjumper
2. See https://github.com/linhd-postdata/spacy-affixes
3. See http://postdata.uned.es:5000/ui/
4. See https:// educalingo.com, https://dirae.es/, and https:// www.fundeu.es/
5. See https://github.com/linhd- postdata/edfu
6. Other corpora, such as DISCO [13] or DraCor [14], were unsuitable for the task since their metrical information was nonexistent or completely computer generated.
7. This distinction only makes sense in Agirrezabal's approach since his neural networks were built independently to accommodate per line or per syllable inputs

Tables for this paper were created using the next Notebook: https://github.com/linhd-postdata/poetrylab-dh2020

Bibliography

[1] Francis, Norbert. Bilingual and Multicultural Perspectives on Poetry, Music, and Narrative: The Science of Art. Lexington Books, 2017.

[2] Poetry Standardization and Linked Open Data (POSTDATA) Project. European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (ERC-2015-STG-679528). http://postdata.linhd.uned.es/ Accessed October 15th, 2019.

[3] Navarro Tomás, Tomás. "Métrica española." Reseña histórica y descriptiva 50. Madrid (1991).

[4] Bello, Andrés. Principios de la ortología y métrica de la lengua castellana, Santiago de Chile, Imprenta de La Opinión, 1835.

[5] Domínguez Caparrós, José. "Métrica española." Madrid, Síntesis (2014).

[6] Caparrós, José Domínguez. Contribución a la historia de las teorías métricas en los siglos XVIII y XIX . Vol. 92. Editorial CSIC-CSIC Press, 1975.

[7] Agirrezabal, Manex, et al. "ZeuScansion: a tool for scansion of English poetry." Journal of Language Modelling 4 (2016).

[8] Navarro-Colorado, Borja. "A metrical scansion system for fixed-metre Spanish poetry." Digital Scholarship in the Humanities 33.1 (2017): 112-127.

[9] Gervás, Pablo. "A logic programming application for the analysis of Spanish verse." International Conference on Computational Logic. Springer, Berlin, Heidelberg, 2000.

[10] Agirrezabal, Manex, Iñaki Alegria, and Mans Hulden. "A Comparison of Feature-Based and Neural Scansion of Poetry." arXiv preprint arXiv:1711.00938 (2017).

[11] Fernández-Carvajal, Francisco. "Antología de textos." Palabra (2003).

[12] Hartman, Charles O. (2005). The Scandroid 1.1.

[13] Ruiz Fabo, Pablo, et al. "The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings." (2018).

[14] Fischer, Frank, Peer Trilcke, and Boris Orekhov. (2019). Drama Corpora Project (DraCor). https://dracor.org/ Accessed October 15th, 2019.

[15] OpenAPI Specification Version 3.0.2. https://swagger.io/specification/ Accessed October 15th, 2019.

[16] Matthew Honnibal and Ines Montani. 2018. spaCy 2: Natural language understanding with Bloom embeddings,convolutional neural networks and incremental parsing. To appear.

[17] Díez Platas, María Luisa, et al. "Hispanic Medieval Tagger (HisMeTag): una aplicación web para el etiquetado de entidades en textos medievales." Presented at Humanidades Digitales Hispánicas 2017, Málaga, Spain.

[18] Fabo, Pablo Ruiz, Clara Martínez Cantón, and Thierry Poibeau. "Distant Rhythm: Automatic Enjambment Detection on Four Centuries of Spanish Sonnets." Digital Humanities 2017. Mexico City, Mexico.

[19] Taulé, Mariona, Maria Antònia Martí, and Marta Recasens. "AnCora: Multilevel Annotated Corpora for Catalan and Spanish." Lrec. 2008.

[20] Padró, Lluís, and Evgeny Stanilovsky. "Freeling 3.0: Towards wider multilinguality." LREC2012. 2012.

[21] Navarro-Colorado, Borja, María Ribes Lafoz and Noelia Sánchez (2016) "Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation", Proceedings of the 10th edition of the Language Resources and Evaluation Conference, 23-28 May 2016, Portorož (Slovenia).

[22] Quilis, Antonio. Estructura del encabalgamiento en la métrica española. Consejo Superior de Investigaciones Científicas, patronato Menéndez y Pelayo, Instituto Miguel de Cervantes. 1964.

[23] Spang, Kurt. Ritmo y versificación: teoría y práctica del análisis métrico y rítmico. Editum, 1983.

[24] Navarro Tomás, Tomás. "Métrica española." Reseña histórica y descriptiva 50 (1991).

[25] Mestre, Antonio Ríos. La transcripcion fonetica automatica del diccionario electronico de formas simples flexivas del español: un estudio fonologico en el lexico. Diss. Universitat Autònoma de Barcelona, 1998.

[26] De la Rosa, Javier, Álvaro Pérez, Salvador Ros, Elena González-Blanco. "Fast and Accurate Syllabification and Scansion of Spanish Poetry." Journal Procesamiento del Lenguaje Natural, 65, 2020.

Linking Time, Space, and Statements in One GIS System: A Use Case of Studying Individuals' Biographies

Yi-Fan Peng
yfpeng@gate.sinica.edu.tw
National Chengchi University, Taiwan Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, Taiwan

Pi-Ling Pai
Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, Taiwan

Chao-Lin Liu
National Chengchi University, Taiwan

Keywords: spatiotemporal analysis, spatialization, biographies studies, geographical information systems
Topics: Asia, English, 20th Century, Contemporary, digital biography, personography, and prosopography, spatial & spatio-temporal analysis, modeling and visualization, Geography and geo-humanities, History

Linking Time, Space, and Statements in One GIS System: A Use Case of Studying Individuals' Biographies

Yi-Fan Peng, Pi-Ling Pai and Chao-Lin Liu[1]

Introduction

Digital humanities research methods have been applied to many research fields in recent years. The content contained in text corpora can be roughly summarized into five Ws[2], especially for biographies. In this research, we focus on spatial analysis and natural language processing (NLP) techniques to find two elements: "Where (Location)" and "When (Time)" in the biographical data. By linking and showing the extracted information via a GIS system, we built a spatiotemporal system for scholars to study historical issues from different perspectives.

Research material and methods

Mr. Chiang Ching-Kuo was a former president of the Republic of China. During his tenure, Taiwan enjoyed a long period of economic growth. The corpus we use is the *Chronological History of Chiang Ching-Kuo* (the *Chronicle*, henceforth) which was compiled by the Chiang Ching-Kuo Foundation.[3] The *Chronicle* records Chiang's life year by year, and the total number of characters exceeds 600,000. The numbers of characters for different periods are shown in Table 1.

Volumes	Number of characters	Volumes	Number of characters
Vol 1 (1910-1944)	51,159	Vol 6 (1974-1976)	88,571
Vol 2 (1945-1951)	71,862	Vol 7 (1977-1979)	92,158
Vol 3 (1952-1958)	76,988	Vol 8 (1980-1983)	80,836
Vol 4 (1959-1966)	54,069	Vol 9 (1984-1988)	53,576
Vol 5 (1967-1973)	101,559		

Table 1. Statistics for each volume

Figure 1 shows the three main steps of data processing, each of which will be explained below.

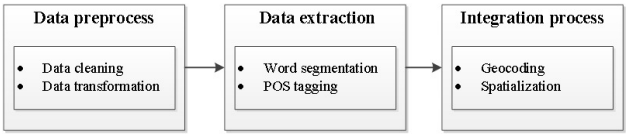


Figure 1. Main steps in data processing

Data preprocess

We converted source texts that are stored as the MS Word files (Figure 2) into pure text files, and we extracted information about the daily activities and their time labels simultaneously.

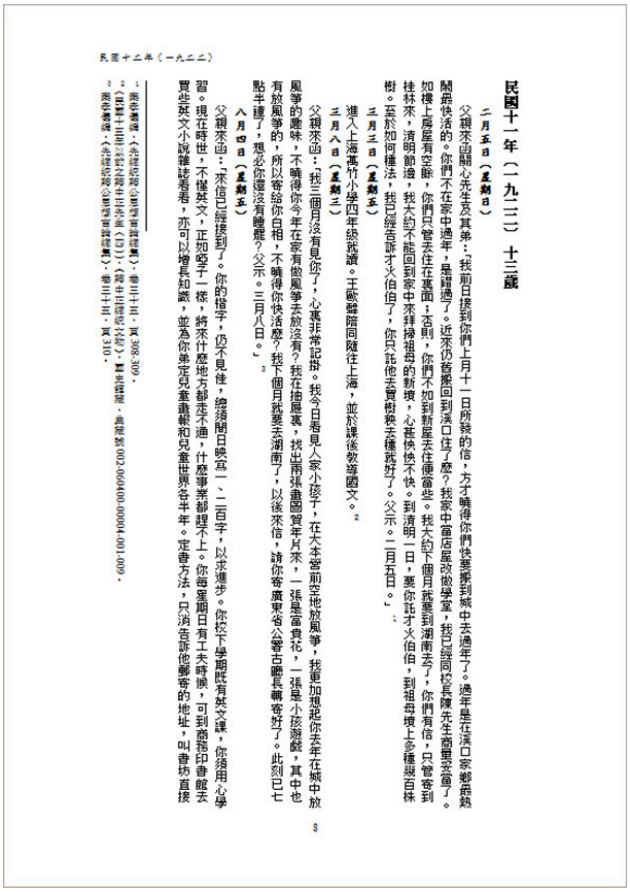


Figure 2. A page of the original material

Data extraction

With the NLP techniques, we can capture the related information for future use. We employ the CKIP[4] tools to segment words and extract the part-of-speech (POS) information. In the corpus, the POS label for the placenames is Nc[5], which is one of the important elements in spatial information. Figure 3 shows the statistics about the quantity of placenames of different lengths.

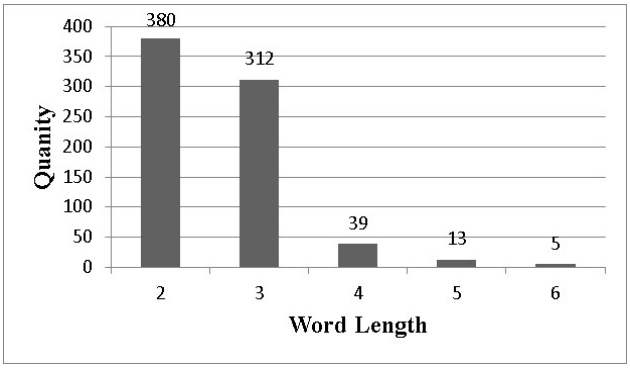


Figure 3. Length distribution of Nc words

Integration process

After extracting the placenames, we geocoded them. Since the *Chronicle* records events after 1910CE, the placenames are modern names. Therefore, we may use the Google Maps API[6] for geocoding. By further combining the temporal information, the placenames can be shown on a temporal scale.

Spatiotemporal System

Next, we construct a WebGIS system that links statements in the *Chronicle* that mentioned the placenames in the GIS subsystem. To visualize the relevance of the corpus to spatial information, the system will show the corresponding original statements in Figure 4 when scholars click on the icons for the placenames in Figure 5.

Figure 4. Original statements in the Chronicle for a chosen placename will be shown for reading

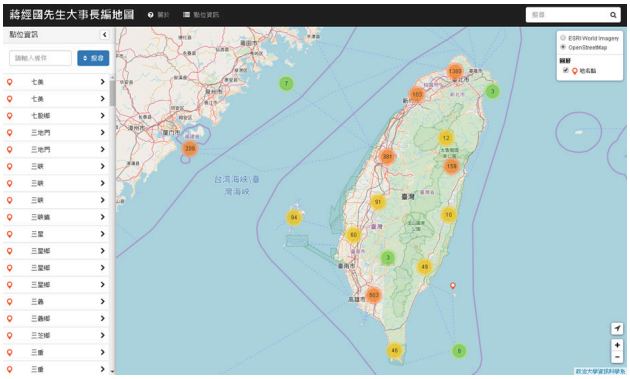


Figure 5. Placenames shown in the WebGIS system and can be clicked to query

Figure 6 shows an example, where the relevant statements were displayed on the left side of the interface. The map shows the spatial locations of the placenames. The time described by the text content is indicated in the timeline at the bottom of the system.

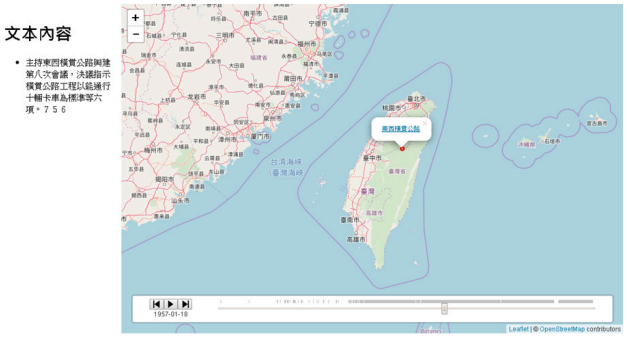


Figure 6. An integration of time, place, and statements

Also, we provide an analysis of the frequency-time relation of placenames. It shows the year in which the placename appeared in the corpus and the frequencies of their appearances. In general, the more frequent a placename is mentioned in a particular period of the text, the more attention it should be paid to the place. Next, we use a practical example to illustrate how to analyze the text content through the system we have built.

The construction of Central Cross-island Highway (?????? is also called ??????) is an important achievement of Mr. Chiang. It is the first highway system that connects the west coast with the east coast of Taiwan. If we want to know the information about the highway in the corpus, we can find out the years of the relevant events by exploring the information of the placenames in our WebGIS and then by exploring the frequency of occurrence of the relevant placenames. There are three aspects to explore this study issue: spatial, temporal, and corpus information.

Spatial information aspect

First of all, we can move the scope of the inspection to a location near the highway through the spatial functions provided by our WebGIS. It will show the placename information related to the location. In this example, we can find information about placenames including "東西橫貫公路"", "天祥I" (tian1 xiang2), "梨山" (li2 shan1), etc., shown in Figure 7. It means that in this spatial range, these placenames are near the highway.

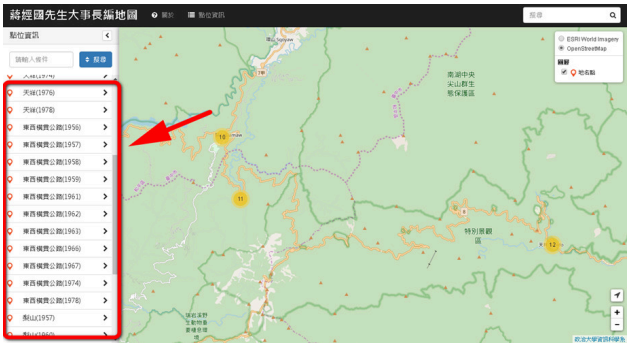


Figure 7. Placenames near the Central Cross-island Highway

Temporal information aspect

By analyzing the frequency and time distribution of the placenames, shown in Figure 8. We may find that the peak frequency of the co-occurrence of these placenames falls between 1956 and 1962.

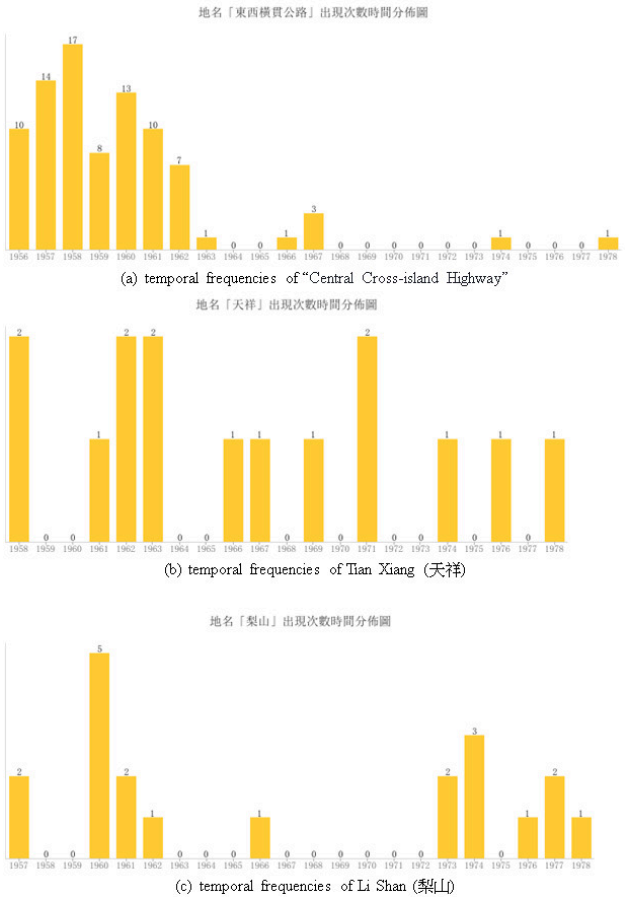


Figure 8. The number of occurrences of placename and its years. (a) is for the placename "東西橫貫公路". (b) is for the placename "天祥". (c) is for the placename "梨山"

Corpus aspect

Corpus aspect

With the spatial and temporal information derived from the previous analysis, we can find out the time period for the road construction and its effects. It can be seen from the text that the period from 1956 to 1962 was the time for the construction of the Central Cross-island Highway. The Tianxiang Scenic Area was established immediately, and the farms in Lishan were gradually developed. With the establishment of traffic routes, these locations along the highway became frequent visits by Mr. Chiang. As such, the researchers can read the statements in the Chronicle for specific times to learn about the relevant events.

The aforementioned example shows that our system allows researchers to conduct a study by first identifying the relevant placesnames via our WebGIS interface. Our system links the placenames to the statements in the Chronicle. Since we have associated the statements with time labels, we could analyze the temporal distributions of the placenames. Naturally, researchers can read the statements on our WebGIS system on a temporal scale as well.

Conclusion

The two elements of time and space play a very important role in the study of biographical data. Through NLP techniques and spatial information methods, we can extract important information from the corpora and build a spatiotemporal system. In this proposal, we illustrate an example of finding the time span for a specific event by analyzing the temporal frequencies of the placenames in the corpus.

Researchers can discover the hidden information among the statements via our system. We believe that showing placenames on a GIS interface and relevant statements about the placenames on a time -related scale can provide helpful hints to complex research work.

[1] Yi-Fan Peng is a doctoral student of National Chengchi University, and a project manager at the Academia Sinica in Taiwan. He can be reached at yfpeng@gate.sinica.edu.tw. Pi-Ling Pai is a post-doctoral researcher at the Academia Sinica in Taiwan, and she can be reached at lingpai@gate.sinica.edu.tw. Chao-Lin Liu is a professor at the National Chengchi University, and can be reached at chaolin@g.nccu.edu.tw.

[2] Five W's means Who, Why, What, Where, and When.

[3] Information about the Chiang Ching-Kuo Foundation is available at http://www.cckf.org/en/.

[4] Ma, W. and Chen, K. (2003). Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff, In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, 17:168-171. DOI: https:// doi.org/10.3115/1119250.1119276

[5] Huang, C. and Chen, K. (1998). Sinica Corpus. DOI:https://doi.org/10.13140/RG.2.1.2357.2963

[6] Google. Developer Guide | Geocoding API [cited 2019 October 12]. Available from: https://developers.google.com/maps/documentation/geocoding/intro.

Towards Open Computer Vision Methods: Visual Analytics of Machine Learning Models in the Digital Humanities

Fabian Offert

offert@ucsb.edu
University of California, Santa Barbara, Germany
Friedrich-Alexander-Universität Erlangen-Nürnberg

Peter Bell

peter.bell@fau.de
Friedrich-Alexander-Universität Erlangen-Nürnberg

Keywords: machine learning, visual analytics, interpretability, computer vision, critical machine vision
Topics: Global, Europe, English, Contemporary, artificial intelligence and machine learning, software development, systems, analysis and methods, Computer science, Media studies

Fabian Offert, Peter Bell

Towards Open Computer Vision Methods: Visual Analytics of Machine Learning Models in the Digital Humanities

Interpretable machine learning, also known as XAI (explainable artificial intelligence), has recently attracted a significant amount of attention in the computer science community and beyond (Lipton 2016, Doshi-Velez and Kim 2017, Gilpin et al. 2018, Olah et. al. 2018). As a relatively new sub-field of machine learning, interpretable machine learning aims to find potential solutions to the inevitable opacity of machine learning models applications by mitigating this opacity through the production of "explanations".

Such a mitigation has become increasingly relevant as machine learning models start to inform socially and politically relevant decision-making processes

(Pasquale 2015, Benjamin 2019, Crawford and Paglen 2019).

From the technical perspective, machine learning models thus become "interpretable" if they, either by design or with the help of external tools, can provide human-understandable explanations for their decisions, predictions, or classifications. In the past five years, various techniques have been developed to produce such explanations, including those targeted specifically at machine learning models operating on images, which are often subsumed under the term visual analytics (Hohman et al. 2018). Explanations in the image domain often function as visualizations, showing and thus explaining how a machine learning model arrives at a certain output.

Beyond these technical results, however, a larger conceptual discussion has emerged in the technical disciplines that "infringes" on the terrain of the digital humanities. It is centered around attempts to find quantitative definitions for concepts that naturally emerge from the problem at hand, such as "interpretation" and "representation", with the help of methods and concepts from disciplines as diverse as psychology, philosophy, and sociology (Kim et. al. 2019, Mittelstadt et. al. 2019, Selbst and Barocas 2018, Ritter et. al. 2017).

Despite the emergence of this cross-disciplinary discussion, however, the interest in interpretable machine learning in the digital humanities community has so far been marginal. Among the few existing studies are Ted Underwood's recent book (2019), which discusses interpretability in the appendix. Arnold and Tilton (2019) briefly examine interpretability in the context of an analysis of the "deepness" of deep learning methods. Finally, Alan Liu's recent work on methods that allow a "close reading of distant reading" are also informed by a discussion of interpretability (Liu 2020). Generally, however, the notion of machine learning models as "impenetrable" black boxes (Arnold and Tilton 2019) persists in the digital humanities community.

Counter to this narrative, in this paper, we argue that interpretable machine learning in general, and techniques from visual analytics in particular, can significantly improve digital humanities research by facilitating a critical machine vision approach to image data. We posit that, generally, interpretable machine learning provides a natural bridge between "close" and "distant" methods. In fact, it provides a critical and technical framework to analyze the "distant" methods themselves. As such, it addresses the question of open data from a meta-methodological point of view: if data is open, the methods which process the data need to be open as well. Importantly, this openness has to go beyond open source. It requires the development of tools that actively mitigate the general opacity of computational methods by making them interpretable, and thus allow them to be both tool and subject of a humanist analysis.

Importantly, the opacity of machine learning models is as much a function of the phenomenological differences between human and machine perception as it is a function of model size (Burrell 2016; Selbst and Barocas 2018; Offert and Bell 2020). This is particularly relevant to digital humanities research on image data, where low-level computer vision methods are often employed to generate high-level hypotheses. For instance, it has been recently shown that deep neural networks often prefer textures over shapes (Geirhos et al. 2019), and generally rely on imperceptible image features (Ilyas et al. 2019). Visual analytics methods can discover such phenomenological biases in regard to specific datasets, and thus lower the risk of misguided higher-level conclusions based on machine predictions while not being subject to a trade-off between model capability and interpretability.

In this paper, we present concrete results for the application of the "feature visualization" visual analytics method (Erhan et. al. 2009, Yosinski et. al. 2015, Olah et. al. 2017) to machine learning models trained on both a toy art historical corpus and the "standard" ImageNet/ILSVRC2012 dataset (Russakovsky et al. 2015). We demonstrate the potential and limitations of the technique, building upon previous work on attribution (Bell and Offert 2020).

Concretely, we show that feature visualization, as a generative approach, offers intuitively interpretable visualizations, which points to an untapped general potential of generative methods like generative adversarial networks (Goodfellow et al. 2014). Feature visualization allows for the discovery of unexpectedly salient image features which can not be detected with attribution methods, and which mirror the phenomenological biases of the employed machine vision systems. In our experiments, for instance, which extend previous work (Offert 2018), it facilitated the discovery of a heavy emphasis on drapery - i.e. a textural feature, as suggested in Geirhos et al. (2019) - in the detection of portraits, counter to the intuitive assumption that the most salient shared feature of portrait imagery would be found in faces.

Applied to the ImageNet/ILSVRC2012 dataset the political dimension of these phenomenological biases becomes evident, where marginal image subjects heavily influence classification. To demonstrate this, we visualized and selected the output neurons for several classes of an InceptionV3 model (Szegedy et. al. 2016) pre-trained on ImageNet/ILSVRC2012 hand-selecting

visualizations that show some non-intuitive properties of the ImageNet dataset.

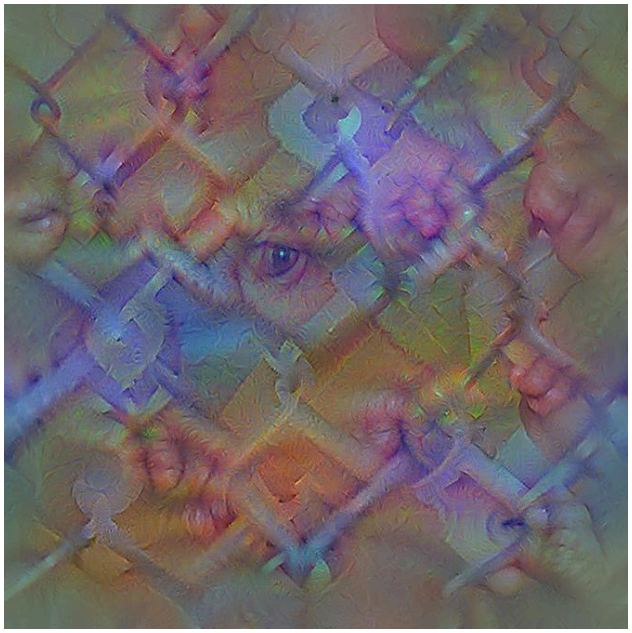


Fig. 1: Feature visualization allows for the discovery of unexpectedly salient image features. Here, a feature visualization of the "fence" class output neuron of an InceptionV3 deep convolutional neural network trained on the ILSVRC2012 dataset shows that, counter to intuition, the detection of fences is influenced by the presence of confined persons, an artifact of the training set. Importantly, this means that the presence of confined persons behind a fence in an image makes that image more fence-like to the classifier.

For instance, for the "fence" class output neuron (fig. 1) we see that the network has not only picked up the general geometric structure of the fence but also the fact that many photos of fences in the original dataset (that was scraped from the Internet) seem to contain people confined behind these fences. This can be verified by analyzing the 1300 images in the dataset class, which indeed show some, but not many scenes of people confined behind fences. Cultural knowledge, more specifically, a concrete representation of cultural knowledge defined by the lense of stock photo databases and hobby photographers, is introduced here into a supposedly objective image classifier. Importantly, this also means that images of people behind fences will appear more fence-like to the classifier. The relevance of this consequence is revealed by a Google reverse image search: for a sample image (fig. 2) from the "fence class", despite the prominence of the person compared to the actual fence, the search produces the Wikipedia entry for "chain-link fencing" (fig. 3), suggesting an unverifiable but likely connection between the Google image search algorithm and ImageNet/ILSVRC2012.

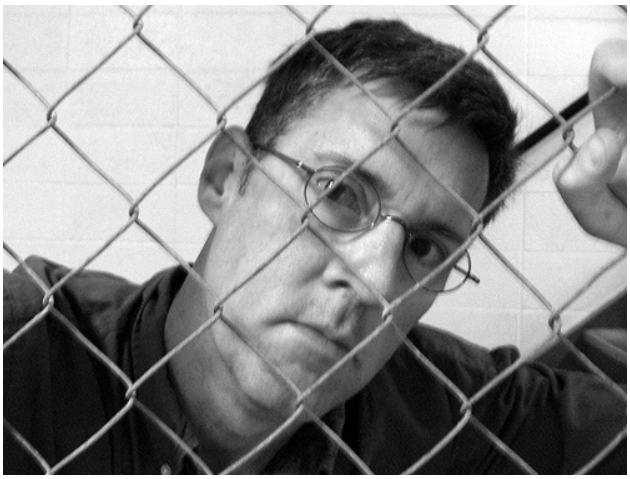


Fig. 2. Sample image from the ILSVRC2012 "chain link fence" class. Note that there are only a few images (between 1% and 5% of the class, depending on what counts as "behind") that show people behind fences.

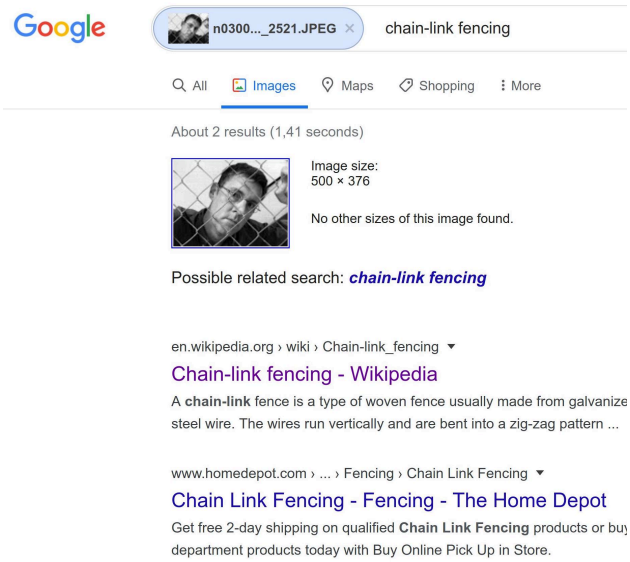


Fig. 3. A Google reverse image search for the selected image, despite the fact that it does not exist on the Internet anymore, and despite the prominence of the person compared to the actual fence, produces the Wikipedia entry for "chain-link fencing", suggesting an unverifiable but likely connection between the Google image search algorithm and ImageNet/ILSVRC2012. A text search for "chain-link fencing" produces no "people behind fences" scenes.

Finally, we propose a concrete analytical pipeline for the visual digital humanities that integrates visual analytics by means of open-source tools and frameworks. For instance, tools like summit (Hohman et al. 2019), lucid (https://github.com/tensorflow/lucid), or tf-explain (https://github.com/sicara/tf-explain/) combine feature visualization and attribution methods to enable a fine-grained analysis of how different neurons in a deep neural network contribute to the classification of an image based on its visual features. We discuss the feasibility of this proposed pipeline in regard to different data domains and computer vision approaches.

We conclude that visual analytics could be integrated into many digital humanities projects, and will positively expand the results of future projects, making digital humanities research not only more rigorous but also increasing its scope.

References

Arnold, Taylor, and Lauren Tilton. 2019. "Depth in Deep Learning: Knowledgeable, Layered, and Impenetrable." https://statsmaths.github.io/pdf/2020-deep-mediations.pdf

Bell, Peter and Fabian Offert. 2020. "Reflections on Connoisseurship and Computer Vision" Journal of Art Historiography, forthcoming.

Benjamin, Ruha. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. John Wiley & Sons.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." Big Data & Society 3 (1).

Crawford, Kate and Trevor Paglen. 2019. "Excavating AI: The Politics of Images in Machine Learning Training Sets." https://www.excavating.ai/

Erhan, Dumitru, Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2009. "Visualizing Higher-Layer Features of a Deep Network." Université de Montréal.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv Preprint arXiv:1702.08608.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. "ImageNet-Trained CNNs Are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness." arXiv Preprint arXiv:1811.12231.

Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning." In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80-89.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." In Advances in Neural Information Processing Systems, 2672-80.

Hohman, Fred Matthew, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. "Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers." IEEE Transactions on Visualization and Computer Graphics.

Hohman, Fred, Haekyu Park, Caleb Robinson, and Duen Horng Chau. 2019. "Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations." arXiv Preprint arXiv:1904.02323.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. "Adversarial Examples Are Not Bugs, They Are Features." arXiv Preprint arXiv:1905.02175.

Kim, Been, Emily Reif, Martin Wattenberg, and Samy Bengio. 2019. "Do Neural Networks Show Gestalt Phenomena? An Exploration of the Law of Closure." arXiv Preprint arXiv:1903.01069.

Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." In 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.

Liu, Alan. 2020. "Humans in the Loop: Humanities Hermeneutics and Machine Learning." DHd2020 Keynote.

Mittelstadt, Brent, Chris Russel, and Sandra Wachter. 2019. "Explaining Explanations in AI." In ACM Conference on Fairness, Accountability, and Transparency (FAT*).

Offert, Fabian. 2018. "Images of Image Machines. Visual Interpretability in Computer Vision for Art." In European Conference on Computer Vision, 710-15.

Offert, Fabian and Peter Bell. 2020. "Perceptual Bias and Technical Meta-Images. Critical Machine Vision as a Humanities Challenge". AI & Society, forthcoming

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. 2017. "Feature Visualization." Distill.

Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. "The Building Blocks of Interpretability." Distill.

Pasquale, Frank. 2015. The Black Box Society. The Secret Algorithms That Control Money and Information. Harvard University Press.

Ritter, Samuel, David G. T. Barrett, Adam Santoro, and Matt M Botvinick. 2017. "Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study." arXiv Preprint arXiv:1706.08606.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. "Imagenet Large Scale Visual Recognition Challenge." International Journal of Computer Vision 115 (3), 211-52.

Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines." Fordham Law Review 87.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. "Rethinking the Inception Architecture for Computer Vision." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818-2826.

Underwood, Ted. 2019. Distant Horizons: Digital Evidence and Literary Change. University of Chicago Press.

Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. "Understanding Neural Networks through Deep Visualization." arXiv Preprint arXiv:1506.06579.

AdA Filmontology – a machine-readable Film Analysis Vocabulary for Video Annotation

Jan-Hendrik Bakels
Jan.Bakels@fu-berlin.de
Freie Universität Berlin, Germany

Thomas Scherer
thomas.scherer@fu-berlin.de
Freie Universität Berlin, Germany

Jasper Stratil
jasper.stratil@fu-berlin.de
Freie Universität Berlin, Germany

Henning Agt-Rickauer
Henning.Agt-Rickauer@hpi.de
Hasso-Plattner-Institute, University of Potsdam, Germany

Keywords: film ontology, time-based arts, linked open data, semantic web, video annotation

Topics: Europe, English, Contemporary, annotation structures, systems, and methods, linked (open) data, Film and cinema arts studies, Media studies

"A film is difficult to explain because it is easy to understand" – this famous quote by Christian Metz (1991, 69) sums up a key challenge of film analysis. The complexity of audio-visual, time-based images, the theoretical premises of film analysis, as well as the overlaps in terminology across different approaches make comparable, systematic film analyses a challenging task. Nevertheless, the possibilities offered by the semantic web as well as linked and open data could change the way in which scholars analyse audio-visual images and relate their findings with each other. The research project 'Audio-Visual Rhetorics of Affect' (https://www.ada.cinepoetics.fu-berlin.de/en/index.html – a collaboration of film studies at Freie Universität Berlin and computational sciences at Hasso Plattner Institute, Potsdam) systematically defined film-analytical concepts and their relationships with each other to make them accessible as a semantic ontology for the structuring of film-analytical data. One of our main goals was to provide an open and expandable ontology for researchers concerned with the detailed analysis of the aesthetics of audio-visual images – not just for the application but also as a structural blueprint for further extensions and alternative modes of description. By means of a) providing an analytical framework for a vast variety of research concerned with moving images (from cinema to TV news and web videos), b) open publication and c) a general openness towards extension and progression, our ontology is designed to serve as an intersection between different academic disciplines and traditions.

This paper focuses on one of the fundamentals of our film-analytical work – the AdA filmontology (http://ada.filmontology.org/) – and discusses some of its implications. The methodological challenge to map reconstructions of film-viewing experience (see Müller/Kappelhoff 2018) within a digital framework resulted in turning the focus on three basic affordances: 1) Creating a vocabulary that is both grounded in a methodological film-analytical consensus as broad as possible while being at the same time applicable with regard to our specific theoretical framework on the aesthetic experience of audio-visual movement-images. 2) Setting up a mode of description that is defined, operationalized and condensed to a degree that allows for the joint annotation of audio-visual corpora by different analysts as well as the (semi-)automatic generation and visualisation of annotations. And that is 3) explicit enough in its definitions to allow researchers

coming from different theoretical backgrounds to relate their approach critically to the analytical data or to adapt our analytical vocabulary to different theoretical frameworks (e.g. neoformalism, narratology or cognitive film theory) by selecting existing and adding new description levels. While the filmontology is currently used to inform qualitative studies on movement patterns, it was published under CC BY-SA 3.0 so that the ontology itself as well as all annotations can be used also in other contexts such as statistical evaluation or citizen science approaches.

In order to make the AdA filmontology usable in film-analytical research, we have participated in developing a video annotation software (https://www.advene.org/) – ADVENE – that enables the authoring and publishing of ontology-based video annotations as Linked Open Data by film scholars without programming skills. As a result, we have developed an environment and a set of tools that on the one hand allow for domain experts to directly participate in the development of a semantic vocabulary through familiar spreadsheet tools, and on the other hand enable the automated generation of a machine-readable OWL (W3C Web Ontology Language) ontology from the provided definitions with the help of converters and mappings [Agt-Rickauer et al. 2018]. These efforts resulted in a systematic vocabulary consisting of more than 400 individual values, assigned to more than 80 annotation types that are organized across 8 general levels of description such as Acoustics, Montage, Image Composition or Camera. For example the level Camera encompasses the types Camera Movement Unit, Camera Movement Type, Camera Movement Speed, Camera Movement Direction, Camera Angle, Camera Angle Canted, Camera Angle Vertical Positioning, Lens, Depth Of Field, Defocus, Recording/Playback Speed (see Figure 01).

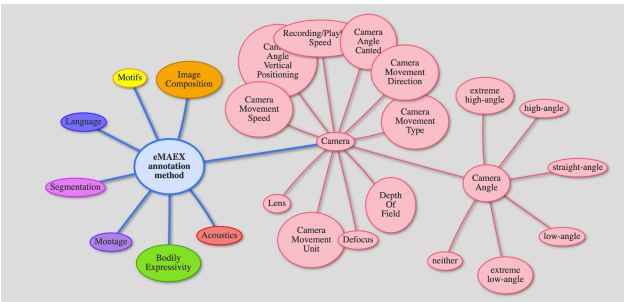


Figure 01: Visualization of the AdA filmontology in Ontoviz.

All values, types, and levels are defined according to the ontology's data model that enables the different kinds of annotations required for in-depth film analysis, including the contrastive or evolving relation between values. The type Camera Angle for instance is defined as: "Perceived vertical angle of (camera) vision. This annotation type provides a scale for camera angles from extreme high angle to extreme low angle." (http://ada.filmontology.org/resource/2020/03/17/AnnotationType/CameraAngle) This scale includes the defined annotation values (extreme high-angle, high-angle, straight-angle, low-angle, extreme low-angle, neither) as well as a 'syntax-element' [TO] allowing for the annotation of an evolving relation (see Figure 02). The definitions of all levels, types and values are published as well at ada.filmontology.org.

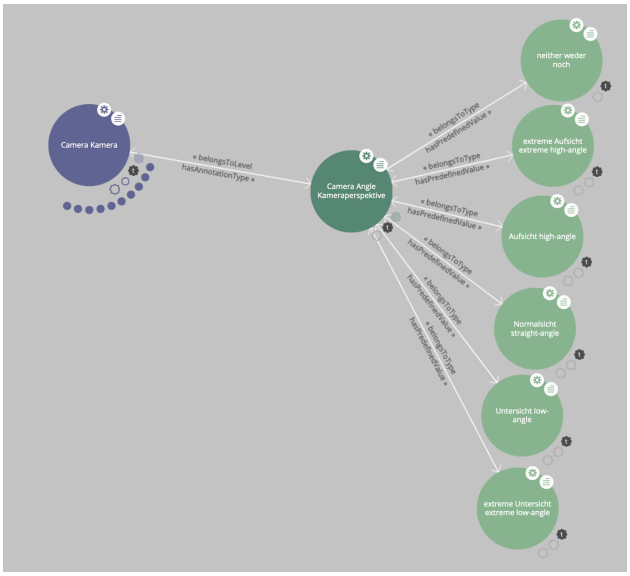


Figure 02: Lodlive visualization of level-type-value-hierarchy for camera angles.

The vocabulary and the definitions are so far bilingual (English and German) with the possibility of adding further languages – thus allowing researchers to interconnect research data across language barriers. In our presentation, we will briefly present a selection of annotation types and values to exemplify the structural and definitional traits of AdA Filmontology that differentiate it from other film-analytical vocabularies, such as Abel Thematic List [Abel 2015] or the Movie Ontology (MO) [Bouza 2010], but also point out similarities to ontologies from other disciplines analyzing time-based media and aesthetic experience (e.g. musicology [Freedman 2019]). Furthermore research projects like Digital Formalism [Heftberger 2018], the Distant Viewing Lab [Arnold/Tilton 2019] as well as the ERC Advanced Grant FilmColors [Flückiger 2017] are important points of reference for our work

regarding the computational analysis of audiovisual aesthetics at the intersection of Film Studies and Digital Humanities with a perspective of exploring the possibilities for further interoperability between approaches.

The AdA filmontology and exemplary annotation datasets using the systematic vocabulary (from feature films, documentaries, TV news and web videos concerning the global financial crisis after 2007) are published regularly under creative commons license. In addition, different ways of visualizing annotations as a 'graphical score' (https://olivieraubert.net/hpi/timeline.html) of audio-visual dynamics (http://ada.filmontology.org/explorer/ ; see Figure 03) and querying these semantically structured annotations (see Figure 04) are developed and published online (https://github.com/ProjectAdA/public) in order to make the results available to and readable for other film scholars as well as researchers from other domains.

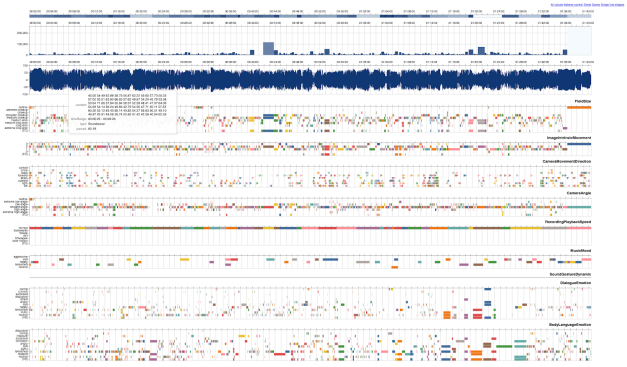


Fig. 03: Several annotation types of a whole feature film in an adjustable and zoomable 'graphical score' generated with Advene.

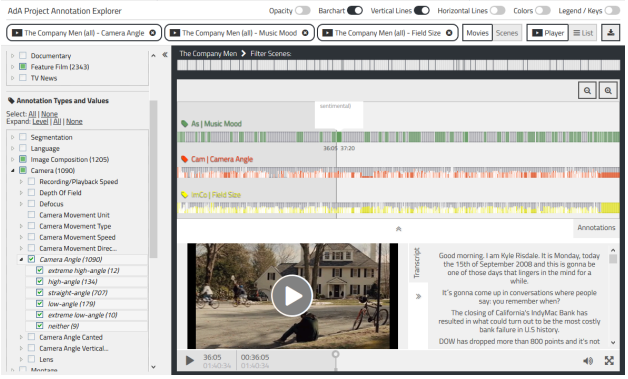


Fig. 04: Querying annotation data in the annotation explorer web application.

References:

[Abel 2015] Abel, Richard (2015) Abel Thematic List V2. onomy.org.

[Agt-Rickauer et al. 2018] Agt-Rickauer, Henning; Aubert, Olivier; Hentschel, Christian; Sack, Harald (2018) Authoring and Publishing Linked Open Film-Analytical Data. EKAW Posters and Demos 2018.

[Arnold/Tilton 2019] Arnold, Taylor; Tilton, Lauren (2019) Distant viewing: analyzing large visual corpora. In: Digital Scholarship in the Humanities, 34, p. i3-i16.

[Bouza 2010] Bouza, Amancio (2010) MO "the Movie Ontology". Movieontology.org.

[Flückiger 2017] Flückiger, Barbara (2017) A Digital Humanities Approach to Film Colors. In: The Moving Image, 17,2, p. 71–94.

[Freedman 2019] Freedman, Richard; Fiala, David; Walter, Micah (2019) The Quotable Musical Text in a Digital Age: Modeling Complexity in the Renaissance and Today. Conference Paper at DH2019 in Utrecht, NL.

[Heftberger 2018] Heftberger, Adelheid (2018) Digital Humanities and Film Studies. Visualising Dziga Vertov's Work. Cham, Springer.

[Metz 1991] Metz, Christian (1991) Film Language. A Semiotics of the Cinema. Chicago, The University of Chicago Press.

[Müller 2018] Müller, Cornelia; Kappelhoff, Hermann (2018) Cinematic Metaphor: Experience – Affectivity – Temporality. Berlin/Boston, De Gruyter.

Topics, buckets, and psychiatry. On the collective creation of a corpus exploration tool

Maria Biryukov
barmaliska@gmail.com
University of Luxembourg, Luxembourg

Roman Kalyakin
roman.kalyakin@ext.uni.lu
University of Luxembourg, Luxembourg

Eva Andersen
eva.andersen@uni.lu
University of Luxembourg, Luxembourg

Lars Wieneke
lars.wieneke@uni.lu

University of Luxembourg, Luxembourg

Keywords: topic modelling, natural language processing, history of psychiatry, cooperation
Topics: Europe, English, 19th Century, 20th Century, digital research infrastructures development and analysis, information retrieval and querying algorithms and methods, History of science, Humanities computing

Topics, buckets and psychiatry. On the collective creation of a corpus exploration tool

Maria Biryukov, Roman Kalyakin, Eva Andersen, Lars Wieneke

Introduction

As described elsewhere (Biryukov et al. 2019) we conducted a probing exercise that brought together a historian and a computer scientist to cooperatively explore a research corpus. With the idea to find and trace the evolution of psychiatric knowledge, the team analysed the *British Journal of Psychiatry*[1] over a timespan of more than 70 years, consisting of 51k pages and covering a wide range of content types such as articles, book reviews, asylum reports, membership lists, etc. A particular challenge was the lack of coherence in the corpus: no logical units such as article boundaries and mediocre OCR quality restricted the direct application of topic modelling and other NLP techniques to a subset of articles through existing tools like MALLET.

The team chose to tackle this challenge by applying topic modelling in order to give structure to the corpus and detect relevant segments for historical analysis. After splitting the corpus into individual pages and applying other preprocessing steps, the team used Non-negative Matrix Factorization (Wang et al. 2012) to decompose the corpus into topics per year, assign words to the topics and to score the individual documents/pages.[2] We then evaluated the optimal number of topics using human evaluation as well as a Word2Vec corpus model (Mikolov et al. 2013) for topic coherence calculation. To enable a diachronic perspective on the corpus we applied dynamic topic modelling to the annual topics. Close inspection of the outcomes made it apparent to the historian that the approach enabled the discovery of weak signals and hidden information that were not easily accessible through conventional approaches.

However, we came to the understanding that a more in-depth historical analysis required a higher degree of autonomy for the historian, as well as facilities allowing multiple perspectives on the data and better ways to present and interact with it. This insight and the potential for a generalisable application convinced us to advance from a probing exercise to building a tool that empowers historians to query, review and explore datasets with a high degree of autonomy and direct access that doesn't require additional mediation.

Building an interface for autonomy and empowerment

Due to the conceptual similarity and availability, we decided to adapt HistoGraph[3] a tool initially designed for the graph based exploration of multimedia collections (Novak et al. 2014; Wieneke et al. 2015; Düring et al. 2015). One of the decisive features of HistoGraph and a strong rationale for integrating our corpus in the tool is found in its drill-down capabilities: all analytical perspectives always lead back to the sources, enabling the researcher to confirm or refute the results of the computational analysis on a case by case basis.

We first performed a named entity extraction task using a combination of Flair's 'ner-fast' model for entity recognition and Google Knowledge Graph for disambiguation.[4] Because of mediocre OCR quality and the historical nature of the corpus, we decided to boost the quality of the entity extraction by aligning the entities with a specialised dataset of historical actors in the domain of psychiatry using a basic string comparison approach.[5] While we initially hoped that the process would yield interpretable network structures, a critical review of the results showed that generic NED models are by design too "contemporary" to successfully identify historical actors and produce meaningful results - an issue we aim to address in the future.

To enable a meaningful perspective on the topic modelling described above, we decided to integrate a new view in histograph that shows a full perspective on the corpus with the ability to "zoom in" on selected parts. In this new view, pages are displayed in sequential order and aggregated in "buckets" of equal size, containing a different number of pages depending on the chosen zoom level. Buckets are displayed as clickable columns in the interface, while topics are displayed in the form of an aggregated heat map of topic distributions across the corpus (*Figure 1*). While working with the historian it became apparent that additional layers would facilitate the corpus exploration, e.g. through adding full-text search capability that presents an aggregated view of text hits across the corpus within the individual buckets.

Collective development and exploration

Of vital importance during our experiments was the constant feedback between the team members. Our discussions led to the insight that more analytical lenses are necessary to better explore the corpus because complex research questions can often not

be captured by a single analytical perspective. In consequence, different layers were added such as the ability to map the results of a keyword search across the corpus. In consequence the pipeline allowed us to discover more relevant content. Our preliminary evaluation compared existing research on the same corpus with our results and we were able to find the same sources as identified by analogue ToC analysis as well as to identify relevant content in other sections that were not accessible through the ToC. For one particular research question - the discussion of the causes of general paralysis in the psychiatric community - this approach proved to be a powerful exploration tool. The extent to which this approach would work on other research questions needs to be explored further and in more detail but looks quite promising.

Brainstorming on this challenge led to the development of an "auto-complete query" feature that builds on the distribution of words in the corpus to refine queries. Initially we used the Word2Vec background model to expand seed terms specified by the researcher. The resulting query was used in the full-text search mode. However, it became apparent that a full-text search based on the constructed query would, from a historians' point of view, lead to a too restricted number of results leaving out promising source material. Looking at this challenge from a computer science perspective, we noticed that the constructed query can be regarded as a vector in the word2vec model that allows to identify documents that are closely related to it, effectively enabling a custom topic for the exploration of the corpus. In close cooperation the team reviewed the feasibility of the approach and ultimately integrated the feature to allow for a more specific yet flexible analysis.

Conclusion and future work

The project started out as a highly dynamic, small scale exploration in the use of topic modelling for structuring an unstructured corpus. Based on the results of this exercise our work was shaped by ongoing discussions between the team members that lead to a co-construction process which not only yielded a new and - as we hope - generalisable tool for the exploration of unstructured or semi-structured corpora, but also shaped the perspectives of the individual team members: by discussing the different approaches used to develop a meaningful interface it allowed the historian an insight into the technology that was used and vice versa enriched the perspective of the computer science colleagues by gaining a deeper understanding of how historians work.

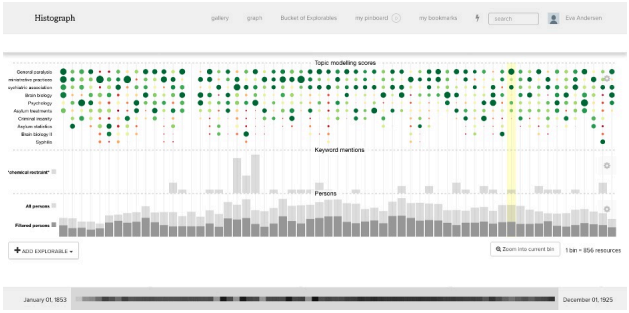


Figure 1. Example of the histogrpah interface displaying a diverse range of topics as well as keyword and person mentions.

Funding

The work of Eva Andersen has been supported by the Luxembourg National Research Fund (FNR) (10929115).

Footnotes

[1] See https://www.rcpsych.ac.uk/about-us/library-and-archives/our-history for details. The association was created as the "Association of Medical Officers of Asylums and Hospitals for the Insane" in 1841, changed its name to "Royal Medico Psychological Association" in 1926 and took its current name in 1971.

[2] We use dynamic topic modelling implementation available at https://github.com/derekgreene/dynamic-nmf for the yearly and dynamic topic modelling, as well as for the Word2Vec-based coherence evaluation.

[3] See http://histograph.eu/ and https://github.com/C2DH/histograph for a docker container, setup instructions and a sample dataset,

[4] We used the Flair 'ner-fast' model trained on the Conll-03 dataset. See https://github.com/zalandoresearch/flair/blob/master/resources/docs/TUTORIAL_2_TAGGING.md#fast-english-models; Google Knowledge Graph see https://developers.google.com/knowledge-graph

[5] The database is used in the 2TBI and TIC Collaborative project, and powered by Nodegoat. Originally it only contained information about Social Reform. See https://www.clariah.nl/projecten/research-pilots/2tbi

Illuminating Past Labor: Making Transformation Processes of Historical Documents Visible

Tomas Vancisin
tv8@st-andrews.ac.uk
University of St Andrews, United Kingdom

Mary Orr
University of St Andrews, United Kingdom

Uta Hinrichs
University of St Andrews, United Kingdom

Keywords: Information Visualization, Data Transformation, Process Visualization, Historical University Records
Topics: Global, Europe, English, 18th Century, 19th Century, data, object, and artefact preservation, spatial & spatio-temporal analysis, modeling and visualization, Design studies, Humanities computing

Historical documents often undergo transformations–transcription, annotation, re-formatting, and digitization–as part of preservation or humanistic inquiry. These processes modify documents' content, artifactual form, and structure, which subsequently influences the ways in which they are read, explored, and interpreted. Transparency of such processes ensures proper attribution of curators' labor, promotes greater inclusiveness, and enables a more holistic and critical interpretation of historical records. But how to engage with and make visible these transformation processes? In this paper, we begin to address this question through a visualization case study based on an exemplary collection of biographical student records from the University of St Andrews (Scotland) that date back to the 18th century. We present – illustrated through this case study - a methodology based on visual (re)-interpretations of historical records over time which, we believe, is relevant to a wide range of information collections within and beyond humanities research.

Biographical Student Records

The handwritten student records (1747-1897) collected by the University of St Andrews originally included students' name, age, church affiliation and birthplace (Fig. 1.1), but over the course of two centuries, they have undergone a variety of transformations (Fig. 1). From 1888 to 1905 the records were transcribed by the University archivist, James Maitland-Anderson (Maitland Anderson, 1905; Fig. 1.2). Between 1960 and 2004, one of Maitland-Anderson's successors, Dr Robert Smart, revised these transcriptions and expanded the records with students' parental lineage, courses taken, and floruit, drawing from a large variety of sources (Smart, 2004; Fig. 1.3). From 2013 to 2016, a team around Dr Alice Crawford from the University library transformed Smart's work into machine-readable form using Text Encoding Initiative (TEI) (Fig. 1.4) which resulted in a web interface that allows targeted searches. In 2017, we transformed the data from Crawford's work into a relational database which enabled the visualization of the records' content (Vancisin et al., 2018; Fig. 1.5). To better understand the nature and impact of such modifications, we conducted interviews with the archivists, historians, and software engineers who have worked on these transformations. In combination with researching previous work on the records (Maitland Anderson, 1905; Smart, 2004), these interviews helped us identify and characterize four key transformations the records have undergone.

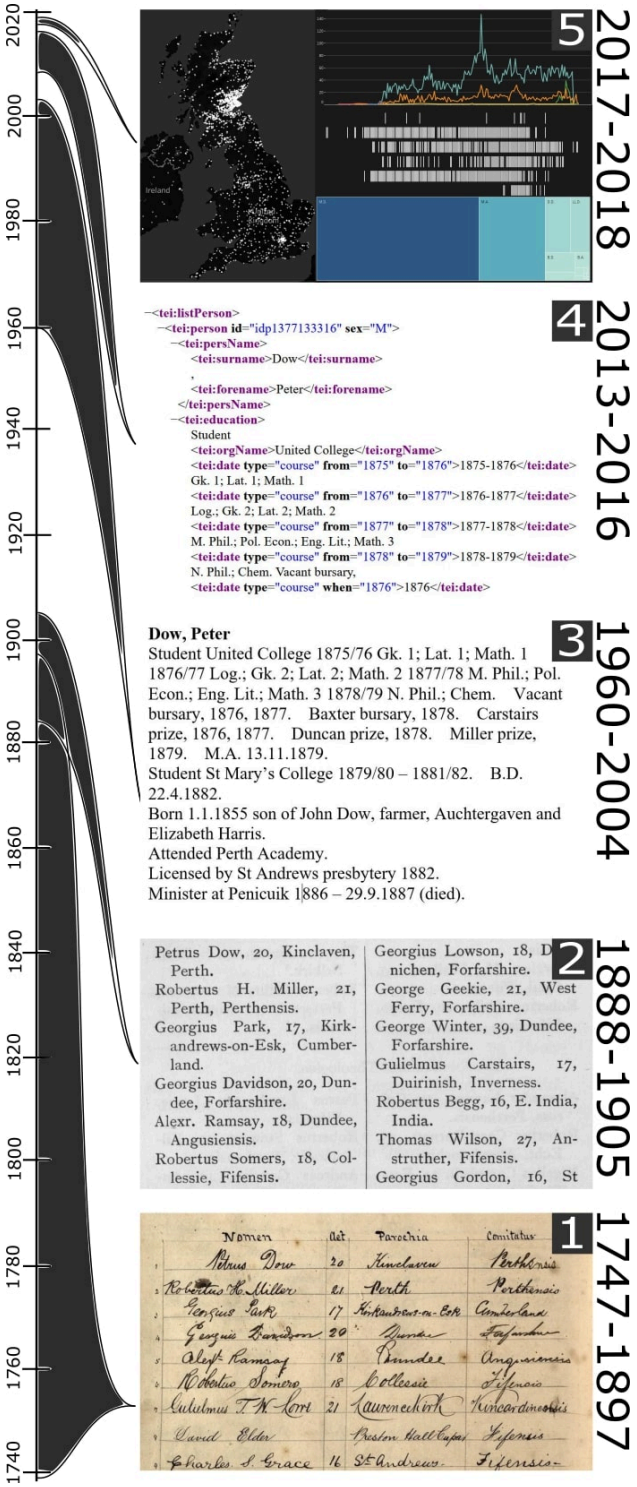


Figure 1: Records' Transformations.

Transcription. Maitland-Anderson and Smart transcribed the original handwritten documents into print form. This process required expertise in paleography and entailed interpretations, for example, of name spellings (e.g., Maitland-Anderson preserved the Latin 'Petrus' Dow; while Smart changed it to 'Peter'; Fig. 1.1-1.3).

Expansion. The records have been expanded by adding information from other archives (Smart), linking records to students' publications (Crawford), and geo-encoding location mentions (Vancisin). These expansions can be considered as interpretations, informed by third-party sources.

Re-Structuring. Maitland-Anderson deliberately preserved the structure of the original records, while Smart transformed them into an alphabetical index, enabling searches by-name but irrevocably removing the records' temporal order. Crawford's tagging imposed a rigid content structure on each record. Our database reconfigures and stores key information (dates, places etc.) in separate but linked tables. This re-structuring allowed new ways of representing and exploring the records, but also introduced additional interpretation layers.

Artifactual Form. Transforming the handwritten records into print enabled easy record parsing, while the structural transformations allowed for new visual and textual representations and interactions. However, the individual human imprint of the handwritten text is lost in these transformations and so are their materiality and visual aesthetics (Forlini et al., 2018). These transformation processes can be considered as re-interpretations of the original records that enable new ways of engagement and representation. However, the fact that the records' transformations are typically invisible is problematic from an ethical (Correll, 2019) and research perspective, because how we represent information fundamentally shapes our interpretation and the questions we ask. Moreover, an unawareness of underlying transformations hampers the holistic interpretation of historical records.

Visualization Opportunities

Historical document visualization mainly focuses on providing access to the content of the collection in its 'final' stage (i.e., Edelstein et al., 2017; Hinrichs et al., 2015; Hyvönen et al., 2017), and our previous work is no exception (Vancisin, 2018; Fig. 2). Work by Hullman & Diakopulos (2011), however, shifts attention to the importance of weaving information about data provenance into visualization. Wrisley (2018) has promoted the idea of Pre-Visualization which argues for visualization prefaces that provide such perspectives in textual form. Péoux & Houllier (2017) introduce a diagrammatic approach to disclose transformation processes. However, visualization-driven approaches that highlight transformation processes and introduce critical and ethical perspectives to the document collection, its metadata, and their visual representations (Correll, 2019, Dörk et al. 2013, D'Ignazio & Klein,

2016) are unexplored. We have started addressing this challenge by investigating how we can portray the records' transformation stages through visualization. Our visualization case study shown in Figure 3 presents one example of how this can be achieved.

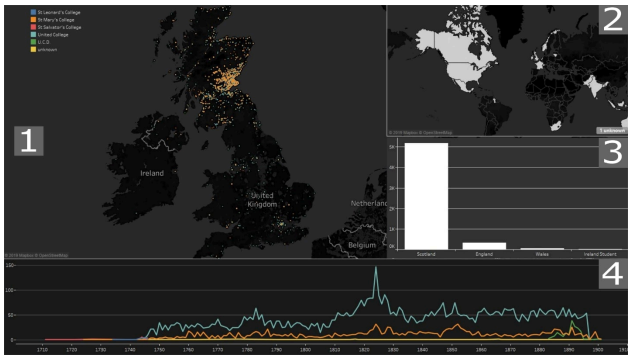


Figure 2: Visualization of the records' content. (1,2) students' birth places within and outside of the UK; (3) distributions of nationalities; (4) student numbers in different colleges over time.

The bar chart at the bottom represents the temporal distribution of students in the original records (Fig. 3.1) while sketchy strokes emphasize the original records' handwritten form. Subsequent transcriptions are depicted in an equivalent bar chart where smooth strokes show the records' transformation into print (Fig. 3.2). The next layer highlights the records' transformation into alphabetical order (reflecting Smart's work) and the content expansion (represented by the bars' width; Fig. 3.3). Crawford's work revoked the records' temporal or alphabetical ordering, so they are depicted as individual squares without any spatial organization (Fig. 3.4; square size corresponds to the amount of information in the record). Our database is shown in a horizontal node-link diagram where rectangles and arcs represent tables and their links (Fig. 3.5). All visualization layers are interactive and interlinked. Hovering over an element in one layer brings up the corresponding records in the record list view (Fig. 3.6) and highlights these in the other visualization layers (Fig. 4, 5 & 6).

Although based on the same data, this type of visualization fundamentally differs from previous approaches in that it enables an exploration of the student records through the lens of their historical context, and through the people involved in their curation and interpretation. Our work, thus, provides a new perspective on visualizing historical documents by illustrating how to allow for their exploration by also considering their history, rather than just their 'final' interpretation. Our visualization of qualitative curatorial changes has to be considered as yet another interpretation of the original records, but we see this approach as an opportunity to make transparent others' and our own interpretations of such collections; it promotes awareness of both the dynamic and interpretative character of historical documents and their visualizations.

Our work combines the categorization of curatorial changes applied to the collection with their visualization, to promote transparency of the (re)-interpretations of the collection over time. Based on this case study, we launch discussion of design principles for visualizations that can make curatorial processes visible, in order to facilitate critical debate that centrally considers key curations of a collection over time (including pre-digital and early digitization), rather than rely only on 'final' data and/or final visualizations which often hide underlying interpretations that led to their assembly.

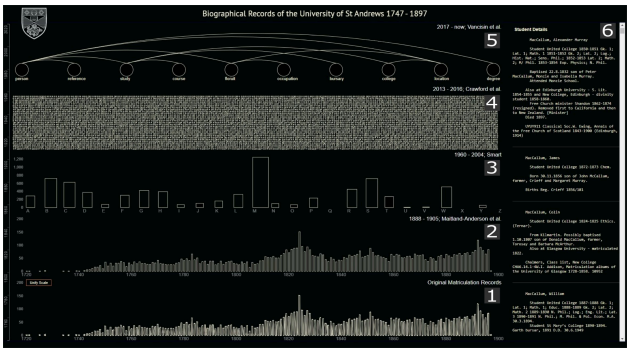


Figure 3: Process Visualization of BRUSA records. (1) Original Records, (2) Transcription, (3) Alphabetical index, (4) TEI files, (5) Database, (6) Student details. [Visualization available at https://tv8.host.cs.st-andrews.ac.uk/stAndrewsStudentBiorecords_contextVis].

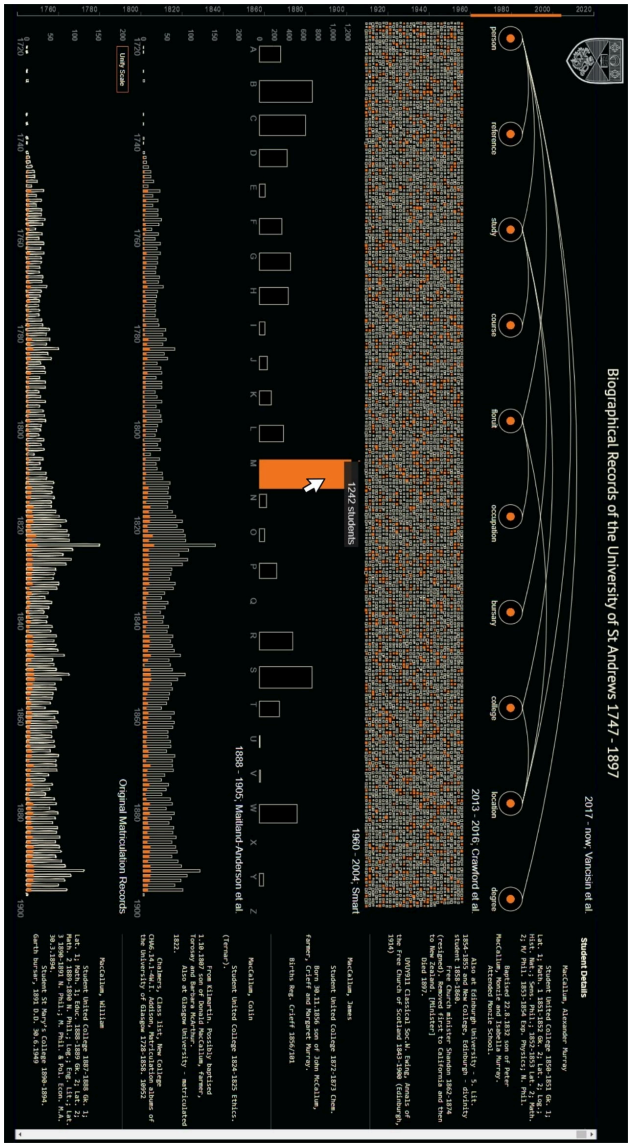


Figure 4: Interconnectedness of the records. Selecting the records starting with the letter "M" highlights these records in all the other visualization layers.

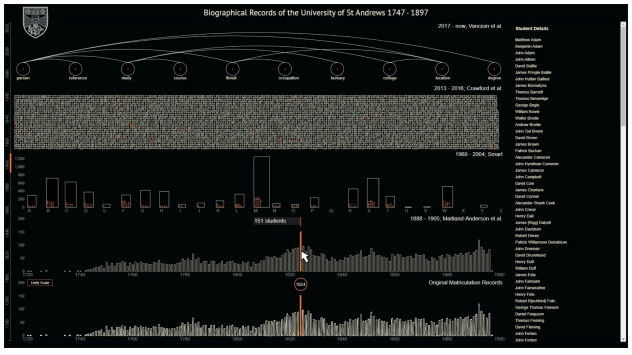


Figure 5: Interconnectedness of the records. Selecting the records from 1824 highlights these records in all the other visualization layers.

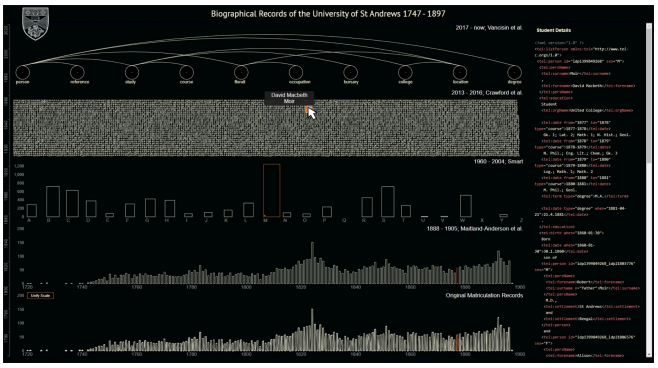


Figure 6: Interconnectedness of the records. Selecting the records from the TEI section highlights these records in all the other visualization layers.

REFERENCES

Correll, M. (2019). Ethical Dimensions of Visualization Research. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19, 1–13. https://doi.org/10.1145/3290605.3300418

D'Ignazio, C., & Klein, L. F. (2016). Feminist data visualization. In Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore. IEEE.

Dörk, M., Feng, P., Collins, C., & Carpendale, S. (2013). Critical InfoVis: exploring the politics of visualization. In CHI'13 Extended Abstracts on Human Factors in Computing Systems (pp. 2189-2198). ACM.

Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., & Coleman, N. (2017). Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project. American Historical Review (p. 25).

Forlini, S., Hinrichs, U., & Brosz, J. (2018). Mining the material archive: Balancing sensate experience and sense-making in digitized print collections. Open Library of Humanities.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., & Coates, C. M. (2015). Trading Consequences: ACase Study of Combining Text Mining and Visualiza-tion to Facilitate Document Exploration. Digital Scholarship in the Humanities (p. fqv046).

Hullman, J., & Diakopoulos, N. (2011). Visualization rhetoric: Framing effects in narrative visualization. IEEE transactions on visualization and computer graphics, 17(12), 2231-2240.

Hyvönen, E., Leskinen, P., Heino, E., Tuominen, J., & Sirola, L. (2017) Reassembling and Enriching the Life Stories in Printed Biographical Registers: Norssi High School Alumni on the Semantic Web. In Language,

Data, and Knowledge (Gracia, J., Bond, F., McCrae, J. P., Buitelaar, P., Chiarcos, C. & Hellmann, S. eds.) vol. 10318 pp.113–119. Cham: Springer International Publishing.

Maitland Anderson, J. (1905). The Matriculation Roll of the University of St Andrews, 1747 – 1897. Edinburgh: Blackwood.

Péoux, G., & Houllier, J. R. (2017). To Visualize Past Communities: A Solution from Contemporary Practices in the Industry for the Digital Humanities. DHQ: Digital Humanities Quarterly, 11(2).

Smart, R. N. (2004). Biographical Register of the University of St Andrews, 1747-1897. St Andrews University Library.

Vancisin, T., Crawford, A., Orr, M. M., & Hinrichs, U. (2018). From people to pixels: Visualizing Historical University Records. in Transimage 2018: Proceedings of the 5th Biennial Transdisciplinary Imaging Conference 2018. pp. 41-57, Transimage 2018, Edinburgh, United Kingdom, 19/04/18. https://doi.org/10.6084/m9.figshare.6104699

Wrisley, D. (2018). Pre-visualization. IEEE 3rd Workshop for Visualization and the Digital Humanities.

Using smart annotations to map the geography of newspapers

Yann Ryan
yann.ryan@gmail.com
The British Library, United Kingdom

Mariona Coll Ardanuy
mcollardanuy@turing.ac.uk
The Alan Turing Institute, United Kingdom

Daniel van Strien
The British Library, United Kingdom

Kasra Hosseini
The Alan Turing Institute, United Kingdom

Kaspar Beelen
The Alan Turing Institute, United Kingdom

James Hetherington
The Alan Turing Institute, United Kingdom

Katherine McDonough
The Alan Turing Institute, United Kingdom Queen Mary University of London

Barbara McGillivray
The Alan Turing Institute, United Kingdom University of Cambridge

Mia Ridge
The British Library, United Kingdom

Olivia Vane
The British Library, United Kingdom

Daniel CS Wilson
The Alan Turing Institute, United Kingdom Queen Mary University of London

Keywords: geographic metadata, toponym resolution, active learning, cultural heritage metadata analysis, spatial analysis
Topics: Europe, English, 19th Century, data, object, and artefact preservation, digital libraries creation, management, and analysis, Geography and geo-humanities, Library & information science

Introduction

Geographic information is a key component in the description of collection objects, and yet its format is often unsuited for use with methods of geographic analysis. Catalogue entries are often inconsistent, in plain text, and without geographic coordinates (much less coordinates linked to authority records). Georesolution of the relevant fields (by matching text strings to citable external resources which do have spatial coordinates) makes catalogue data machine-readable and allows collection exploration that more fully takes the geographic dimension of metadata into account.

Geographic metadata analysis requires a very high-quality resolution of the relevant metadata fields. However, georeferencing by hand is highly time-consuming, and both state-of-the-art georesolution systems and off-the-shelf geocoders achieve highly varying performance depending on the datasets they are applied to (Gritta 2019, DeLozier 2015, Alex 2015). Here, we propose a strategy to resolve place names in metadata that uses an active learning method based on heterogeneous uncertainty sampling (Lewis and Catlett 1994). This method, which we call "smart

annotation" (because it depends on selective human feedback), significantly reduces the number of manual annotations by actively querying the user only for less certain matches. We applied this method to the British Library (BL) newspaper title catalogue and obtained 25,000 high-quality georeferenced records in less than three hours.

The value of geographic newspaper metadata

The BL holds about 37,000 newspaper titles, of which 25,000 are from inside the UK and Ireland. BL newspaper titles are catalogued with geographic metadata in plain text, including country, county, and city. This metadata does not use a controlled vocabulary, and the city and county are not necessarily separated hierarchically. According to Tim Sherratt, the value of digitised newspapers collections is that we can use the 'granularity of newspapers to expose the local, the particular, the personal and the ephemeral' (Sherratt 2015). The nineteenth-century newspaper landscape of the UK was particularly based around local and regional titles (Hobbs 2018). Geographic analysis using computational methods requires that the geographic data is correct: if we are to understand the local and regional differences of our collection items, they must be correctly and systematically resolved to a place.

Method

The "smart annotation" pipeline consists of two steps: an automatic resolution step and an interactive curation step.

The first step follows a traditional toponym resolution process: we preprocess the relevant geographic metadata fields to obtain the main toponym and (potentially) a list of context toponyms that help to disambiguate the location of the main toponym. For example, given the two columns "Annan, Dumfries & Galloway" and "Scotland, UK", we consider "Annan" as the main toponym and "Dumfries", "Galloway", "Scotland", and "UK" as context toponyms. Using a Wikipedia-based gazetteer enhanced with information from Geonames, we select for each toponym (main or context) all potential candidates that might be their true referent, and attempt to resolve the true location of the main toponym taking the following features into account for each candidate:

- Confidence of the candidate referring to the toponym, depending on the source from which it has been obtained (Wikipedia page, Wikipedia redirection, Geonames altnames, etc),
- Relevance of main and context toponym candidates,
- Geographic distance between the different context toponym candidates with respect to the main top onym candidates,

Whether the candidate location is in the country of the collection (UK in the case of the BL newspaper titles collection).

The best candidates are selected through a scoring function that combines the different features to obtain the most likely resolution. We use this scoring function to calculate a confidence score for each main resolved toponym. This confidence score is then passed to the second step, in which results from the resolution process are uploaded to a Jupyter notebook where the annotator can validate each resolution and correct it if required.

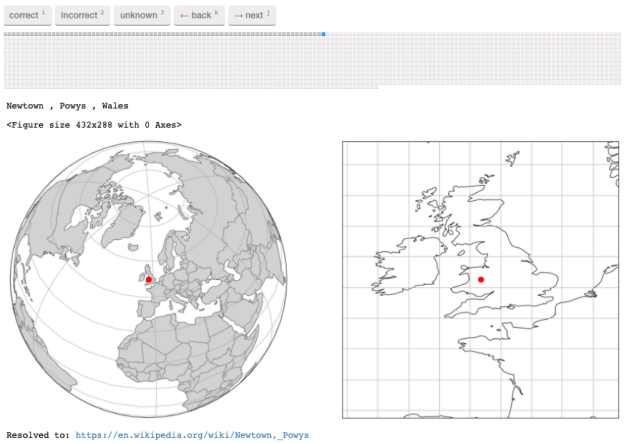


Figure 1: Validation step within a Jupyter notebook cell: 'Newtown, Powys, Wales' is the original metadata text; the automatic resolution is provided in two forms: (1) as global- and regional-scale maps containing the resolved location, and (2) as the link to the Wikipedia page to which the main **toponym** has been resolved. By tagging the resolution as incorrect or unknown, the annotator can then choose any other of the considered candidates or manually input the correct latitude and longitude.

Figure 1, we use the ipyannotate Jupyter Widget for data annotation: https://github.com/ipyannotate/ipyannotate.

The results to be annotated are sorted from low to high confidence, ensuring that those resolutions for which the system was less certain are validated and corrected first (Figure 2).

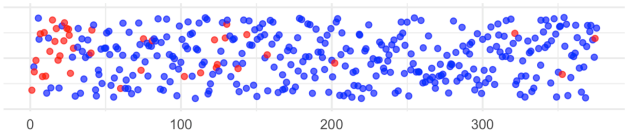


Figure 2: Manually validated annotations sorted by resolution confidence: correct resolutions in blue, incorrect resolutions in red. The x-axis shows the number of annotations completed. The points have been distributed randomly along the y-axis for visibility.

Results

We have applied smart annotation (resolution, validation, and correction) to the ca. 25,000 newspaper titles from UK and Ireland. It took less than three hours of annotations to achieve an accuracy of 97% (allowing for an error margin of 10 miles between the resolved place and the gold standard).

These results can be used to create accurate maps and conduct further geographic and demographic analysis. They help us to understand, for example, the changing focus of the press from national, to regional, to local. This disambiguation method has the potential to have a large impact on the field of newspaper and periodical studies, as high-quality metadata linked to a standard external resource allows us to understand both the collection and the wider newspaper landscape. Ultimately, the goal is to use this information to enhance and enrich existing BL metadata for the benefit of library users, so they can better understand the collection. We demonstrate the value of this technique with a visualisation tool we created using R, Shiny and Leaflet (see Figures 3 and 4).

Resolved geographic metadata can help in understanding the geographic scope of the news collection. It enables the discovery of geographic bias in the digitised portion of a collection and enables collection data to be joined to census data to do demographic analysis. Because newspapers are richly geographical, correctly resolved geographic metadata can aid in the disambiguation of digitised newspaper text. The results presented here are based on the performance of a toponym resolution system (and of a gazetteer) that works well with the kinds of toponyms one can encounter in BL newspaper titles metadata: in English, and mostly focused on Great Britain and Ireland. In the future, we intend to apply this method to other metadata collections.

* Code will be available on the Living with Machines GitHub organization: https://github.com/Living-with-Machines

References

Alex, Beatrice, Kate Byrne, Claire Grover, and Richard Tobin. "Adapting the Edinburgh Geoparser for Historical Geo-referencing." IJHAC: A Journal of Digital Humanities 12, no. 1 (2015).

DeLozier, Grant, Jason Baldridge, and Loretta London. "Gazetteer- Independent Toponym Resolution Using Geographic Word Profiles." In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2382-88. AAAI'15. AAAI Press, 2015.

Gritta, Milan, Mohammad Taher Pilehvar, and Nigel Collier. "A Pragmatic Guide to Geoparsing Evaluation." CoRR abs/1810.12368 (2018). http://arxiv.org/abs/1810.12368

Hobbs, Andrew. A Fleet Street in Every Town: The Provincial Press in England, 1855-1900, 2018.

Lewis, David D., and Jason Catlett. "Heterogeneous Uncertainty Sampling for Supervised Learning." In Proceedings of the Eleventh International Conference on Machine Learning, 148-56. Morgan Kaufmann, 1994.

Sherratt, Tim. "Asking Better Questions: History, Trove and the Risks That Count." In Copyfight, edited by Phillipa McGuinness, 112-24. NewSouth Publishing, 2015.

Historical Newspaper Content Mining: Revisiting the impresso Project's Challenges in Text and Image Processing, Design and Historical Scholarship

Maud Ehrmann
maud.ehrmann@epfl.ch
École polytechnique fédérale de Lausanne (EPFL), Switzerland

Estelle Bunout
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Simon Clematide
Universität Zürich, Switzerland

Marten Düring
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Andreas Fickers
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Roman Kalyakin
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Frédéric Kaplan
École polytechnique fédérale de Lausanne (EPFL), Switzerland

Matteo Romanello
École polytechnique fédérale de Lausanne (EPFL), Switzerland

Paul Schroeder
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Philipp Ströbel
Universität Zürich, Switzerland

Thijs van Beek
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Martin Volk
Universität Zürich, Switzerland

Lars Wieneke
Luxembourg Centre for Contemporary and Digital History (C2DH), Luxembourg

Keywords: Historical newspapers, text mining, interface design, digital source criticism
Topics: Europe, English, 19th Century, 20th Century, Contemporary, Interface design, development, and analysis, text mining and analysis, Computer science, History

impresso. Media Monitoring of the Past is an interdisciplinary research project in which a team of computational linguists, designers and historians collaborate on the datafication of a multilingual corpus of digitised historical newspapers. The primary goals of the project are to improve text mining tools for historical text, to enrich historical newspapers with (semi-) automatically generated data and to integrate such data into historical research workflows by means of a newly developed user interface. In this paper we discuss our efforts to overcome inherent challenges and to integrate text mining and data visualisation applications in general historical research practices which are characterised by search operations as well as the need to create topical collections.

Serving the city: an automatic information extraction for mapping Amsterdam nightlife (1820-1940)

Leon van Wissen
l.vanwissen@uva.nl
University of Amsterdam, The Netherlands

Claartje Rasterhoff
University of Amsterdam, The Netherlands

Thunnis van Oort
University of Amsterdam, The Netherlands

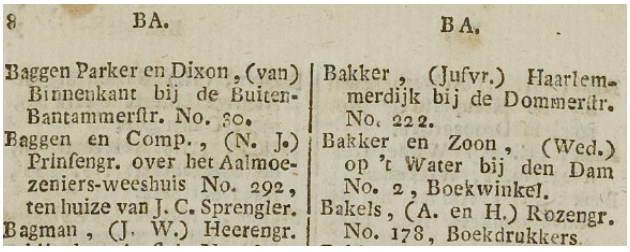
Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, The Netherlands

Keywords: Semi-structured Text, Entity Reconstruction, Nightlife, Urban History, Pleasurescapes
Topics: Europe, English, 19th Century, 20th Century, data modeling, information retrieval and querying algorithms and methods, Cultural studies, History

This paper discusses how automatic information extraction and linked data can facilitate long-term socio-spatial analyses in urban history, using as a case study a long-term spatial analysis of the development of urban nightlife and entertainment industry in Amsterdam. We demonstrate how automatic information extraction of civil and trade registries can corroborate and validate information found in these sources, using duplicate information from the sources themselves as well as other published corpora [1] that contain overlapping information. We present a technique that greatly reduces the need for manual revision of the data. Such a technique is dearly necessary - often research projects resort to hiring people to manually go through all the data or resort to crowdsourcing, which is expensive, slow, and has the problem that the data has to be looked at by multiple non-experts in order to find 'likely correct' data (see e.g. Zheng et al. 2017). Previous efforts to do such a thing have focused on cleaning OCR results by using contextual probabilities or word

frequency probabilities (e.g. Reynaert 2014; Mei et al. 2018; Jatowt and Nguyen 2018; Doush, Alkhateeb and Gharaibeh 2018), at times combined with semantic data (e.g. Woodfield et al 2018), which works on texts that are representative of written vocabulary but is not very apt for lists of names and places. Our technique combines and compares several archival sources that include similar information on, for instance, person and place names. Thereafter, a probability estimation disambiguates between one or multiple occurences of references to what is most likely the same entity. To keep track of and to document this process we model and publish the data in the new roar ontology, [2] which allows us to model archival data, while keeping the provenance chain of decision making and entity disambiguation.

Informed by recent developments in the field of digital history as well as sociological literature on urban nightlife, this paper applies these techniques to address the underrepresentation of small, short-lived, and informal cultural venues in quantitative studies on cultural life. Even though the nineteenth and twentieth century witnessed the rise of nightlife industries as we know them today (Baldwin 2012; Schlör 1998; Nasaw 1999, Erenberg 1981), the organization of small entertainment venues and their effects on the production and consumption of urban cultural life have received little systematic analysis beyond case studies on individuals and single venues, especially in the Dutch context. This is partly due to the large number of yet undigitized archival periodicals such as address books, almanacs and program listings. However, the advent of new methods for digitizing and analyzing documents allowed us to develop a new and less labour intensive technique for uncovering and structuring these valuable sources.


8 BA. BA.

Baggen Parker en Dixon, (van) Binnenkant bij de Buiten-Bantammerſtr. No. 80.
Baggen en Comp., (N. J.) Prinſengr. over het Aalmoezeniers-weeshuis No. 292, ten huize van J. C. Sprengler.
Bagman, (J. W.) Heerengr.
Bakker, (Jufvr.) Haarlemmerdijk bij de Dommerſtr. No. 222.
Bakker en Zoon, (Wed.) op 't Water bij den Dam No. 2, Boekwinkel.
Bakels, (A. en H.) Rozengr. No. 178, Boekdrukkers.

Figure 1 - Address book example (1824). Usual structure is: [name] [(initials/prefix)] [street name] [house number] [occupation] [either telephone number or neighbourhood number]

In this paper we help to redress this issue of underrepresentation by using automatic information extraction and entity disambiguation to systematically trace and analyse the spatial and diachronic distribution of clubs, restaurants and bars in relation to more formal cultural venues such as theatres and concert halls, as well as the rise of new entertainment venues such as cinemas. To reconstruct this urban pleasurescape, we contrast established historiographical narratives with a diverse set of historical sources. We reconstruct the public sphere of food and drink consumption by digitizing address books, the 'yellow pages' of those times (see Figure 1 for an excerpt of such a book), that listed most businesses in Amsterdam, their locations, and proprietors' names. The data in the address books is semi-structured and rather consistent over time, whereby it was possible to programmatically extract information. We extracted all persons that had occupations in the food and drink service industry such as 'koffijhuishouders' [coffee shop holders] and 'tappers' [tavern keepers], but also theatres, cinemas and music venues. Once digitized, an automated comparison of these books significantly reduces the need for manual correction of OCR errors, spelling variation, other artifacts and errors. [3] Other sources, such as the Amsterdam Citizen Registry, are used to validate the automatic extraction from the address books.

We validate the method by gathering as many sources with overlapping information as possible, normalizing its contents slightly (e.g. orthography of occupations and street names) and scoring each bit of information to deduce which data is likely correct and which is not by taking into account source reliability, temporal and geographic distance, and number of sources giving the same and dissenting information. This also provides an order of probability of necessity of manual revision that makes it possible to address the problems in an efficient manner; the data is most quickly improved correcting the certainly wrong information first, and leaving the possibly wrong for later. For address books of subsequent years alone, the match rate is around 50%, which means that half the data does not need checking. For larger gaps between subsequent address books the match rate drops, since the information in the books differs increasingly over time. Adding a second address book from the year preceding the earliest increases the number to 60%. After linking the street names in the books with georeferenced data on Adamlink, we could plot their locations and see the changing patterns over time, demonstrating how patterns of urban expansion impacted on the organisation of urban nightlife in the city of Amsterdam between 1820 and 1940. Finally we fit and publish the data in the roar model, thereby creating transparent linked data reconstructions of historic persona and businesses for others to use.

This systematic digital historical approach has, moreover, the added value of making source bias explicit, as the comparisons show which information present in one is not present in another. The bottom-up approach reveals significant gaps in our current knowledge of urban nightlife. It also shows the value of using multiple sources that overlap in the information they carry in reducing labour and increasing accuracy, and introduces roar as data model for storing this type of information.

[1] See primary sources.

[2] https://w3id.org/roar.

[3] We built on our experiences in the DIGIFIL project, where we developed other techniques to extract data from Dutch film listings over a 50-year period (Kisjes et al. 2019).

Bibliography

Primary sources

Amsterdam City Archive, collection 30274: Adresboeken [address books], inventory 1-96 (coverage 1821-1939). https://archief.amsterdam/inventarissen/overzicht/30274.nl.html

Amsterdam City Archive, collection 30398: Koopmansboekjes [merchant books], inventory 32-37 (coverage 1821-1838). https://archief.amsterdam/inventarissen/overzicht/30398.nl.html

Amsterdam City Archive, Index of the Bevolkingsregisters [citizen registries] (coverage 1851-1893). https://archief.amsterdam/inventarissen/overzicht/5000.nl.html

HisGis, Fryske Akademy, https://hisgis.nl/projecten/amsterdam/

Datasets, tools and ontologies

Adamlink - https://adamlink.nl/

roar, Reconstructions and Observations in Archival Resources - https://w3id.org/roar

Tesseract OCR - https://github.com/tesseract-ocr/tesseract

Secondary sources

Baldwin, P. C. (2012). In the Watches of the Night: life in the nocturnal city, 1820-1930. University of Chicago Press.

Doush, I. A., Alkhateeb, F., & Gharaibeh, A. H. (2018). A novel Arabic OCR post-processing using rule-based and word context techniques. International Journal on Document Analysis and Recognition (IJDAR), 21(1-2), 77-89.

Erenberg, L. A. (1981). Steppin'Out: New York Nightlife and the Transformation of American Culture. University of Chicago Press.

Coustaty, M., Doucet, A., Jatowt, A., & Nguyen, N. V. (2018, November). Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction. In International Conference on Asian Digital Libraries (pp. 278-289). Springer, Cham.

Kisjes, Ivan, K. Beelen, T. van Oort, K. Lotze: Mining the Movie Landscape. Extracting Film Listings from Digital Newspapers. September 2019. Presented at DHBenelux 2019, 11-13 September. Liège, Belgium.

Mei, J., Islam, A., Moh'd, A., Wu, Y., & Milios, E. (2018). Statistical learning for OCR error correction. Information Processing & Management, 54(6), 874-887.

Nasaw, D. (1999). Going out: The rise and fall of public amusements. Harvard University Press.

Reynaert, M. (2014, August). TICCLops: Text-Induced Corpus Clean-up as online processing system. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations (pp. 52-56).

Schlör, J. (1998). Nights in the big city: Paris, Berlin, London 1840-1930. Reaktion Books.

Zheng, Y., Li, G., Li, Y., Shan, C., & Cheng, R. (2017). Truth inference in crowdsourcing: Is the problem solved?. Proceedings of the VLDB Endowment, 10(5), 541-552.

Woodfield, S. N., Seeger, S., Litster, S., Liddle, S. W., Grace, B., & Embley, D. W. (2018, October). Ontological Deep Data Cleaning. In International Conference on Conceptual Modeling (pp. 100-108). Springer, Cham.

All papers are data papers: from open principles to digital methods

Arthur Perret

arthur.perret@u-bordeaux-montaigne.fr
Université Bordeaux Montaigne, laboratoire MICA E3D

Olivier Le Deuff

Université Bordeaux Montaigne, laboratoire MICA E3D

Keywords: documentarity, digital notebooks, FAIR data, Paul Otlet
Topics: Europe, English, North America, 20th Century, Contemporary, data publishing projects, systems, and methods, meta-criticism (reflections on digital humanities and humanities computing), Library & information science

How do we bridge the gap between ambitious global schemes, such as Paul Otlet's "Aims of documentation" (Otlet, 1934) or the FAIR data principles

(Wilkinson et al., 2016), and existing information practices? We describe the theoretical basis and practical steps for a subject-oriented approach to this problem, examining data-related expectations through the lens of documentarity.

In 1934, Belgian bibliographer Paul Otlet published a *Treaty of documentation* in which he outlined the "Aims of documentation":

"Universal as to their purpose; reliable and true; complete; fast; up to date; easy to obtain; collected in advance and ready to be communicated; made available to the greatest number of people." (Otlet, 1934, p. 6)

In 2016, the FAIR principles were published along similar lines:

"To be Findable; to be Accessible; to be Interoperable; to be Reusable." (Wilkinson et al., 2016, p. 4)

They differ in some ways: Otlet viewed the Aims as a whole, with openness as a critical element, while FAIR is modular and not necessarily synonymous with open data. But more importantly, they both describe a plan which is meant to precede and guide implementation. Otlet's Aims are broken down into goals related to the actual "*biblio-technie*" or "*bibliothéconomie*" (Otlet, 1934, pp. 372-375); similarly, each of the 4 components of FAIR is itself divided in 4 sub-components which delve into technical matters (e.g. data vs. metadata). These are actionable steps to be applied in the field, which is where trouble begins.

During and after his time, close collaborators and distant peers alike noted the gap between Otlet's ambitions and what he was able to achieve: Valère Darchambeau commented on "Mr. Otlet's mental audacities, his utopias some would say" (Mundaneum archives, PP P0 462); Suzanne Briet called him ironically "the magus" of documentation. Indeed, he had a major impact on the institutionalization of documentation- the development of Library and information science (LIS) in Europe owes much to section 4 of his *Treaty*- but his work on the relationship between subject and knowledge was largely neglected. The techno-semiotic mediations of information have been far less studied in LIS than human ones; we can arguably trace this back to Otlet's incomplete legacy. Conversely, the implementation of FAIR principles quickly raised the issues of user experience, expectations and metrics:

"FAIRness is aspirational, yet the means of reaching it may be defined by increased adherence to measurable indicators . . . metrics that reflect the expectations of particular communities." (Wilkinson et al., 2017, pp. 1-2)

The interface between person and information seems much thinner for computer-held data than for library books. While this is not actually true (mediations have simply shifted towards human-computer interaction), it means that the feasibility of principles is challenged almost immediately by subjective experience. Data may be FAIR but people may differ: they do not all work on the same data or with the same mindset and therefore have different expectations. This shapes the way we assess data within the framework of documentation and therefore its value to us-its documentarity.

Documentarity is the product of interdisciplinary theoretical work, at the intersection between ontology, documentation and linguistics. The first of these two influences have been studied: documentarity can be seen as a philosophy of evidence based on documentation (Day, 2019) and also as the quantifiable documentary quality of things, with applications to digital documents and data (Perret & Le Deuff, 2019). Here, we examine the third influence: how linguistics contribute to documentarity as an epistemological proposal which at the core focuses on the reception of information. We show that documentarity is linked to several works: Roman Jakobson's "*literaturnost*", which in French ("*littérarité*") (Jakobson, 1977, p. 16) is very close to documentarity ("*documentarité*"); Hans Robert Jauss' adaptation of horizons of expectation ("*Erwartungshorizont*") to literature (Jauss, 1970); the shape of enunciation with Mary-Ann Caws "*architexture*" (Caws, 1981, p. 10) and Roger Laufer's "*scripturation*" (Laufer, 1986, p. 75).

This array of concepts is dense but its purpose is coherent: we draw from the phenomenology of the reading process to make better sense of the way we assess computer-held data. Our methodology is to track the embodiment of thought in technological mediations, especially in writing. The usefulness of such an approach has been described for the study of information as experience (Gorichanaz, 2017). We argue that the way we perceive the documentarity of data is shaped by our horizons of expectation, especially previous experience of genre-based rules which me must establish if we wish to prevent global principles from falling into abstraction as soon as they enter the field.

In this perspective, digital notebooks form a stimulating case study, highly relevant to the conference's theme on open data. They relate to a tradition and to new practices (data science, data papers). We analyze the way data is presented and interacted with in R, Python and Javascript-based notebooks, and we observe a reflexive impact on our perception of documentarity: it allows us to relate more practically to the intellectual framework behind Otlet's "Aims of documentation" and the FAIR principles, which could improve their adoption. Through reproducibility and replicability, the practice of the

notebook informs us on the relationship between data and truth. It also underlines the status of text as the most basic and universal type of data in science: the way text is handled in notebooks (lightweight markup languages, integration of standards, automation) shifts our perception of 'text' to 'textual data'. This is independent from the field of study: we suggest that any research built from plain text can be considered a data paper and that extending "FAIRness" to scientific writing in general would be an epistemological breakthrough in scientific communication.

References

Caws, M. A. (1981). The eye in the text. Princeton University Press. http://archive.org/details/eyeintextessayso0000caws

Day, R. E. (2019). Documentarity: Evidence, Ontology, and Inscription. MIT Press.

Gorichanaz, T. (2017). Auto-hermeneutics: A phenomenological approach to information experience. Library & Information Science Research, 39(1), 1-7. https://doi.org/10.1016/j.lisr.2017.01.001

Jakobson, R. (1977). Huit questions de poétique (T. Todorov, Ed.). Éd. du Seuil.

Jauss, H. R. (1970). Literary History as a Challenge to Literary Theory. New Literary History, 2(1), 7-37. https://doi.org/10.2307/468585

Laufer, R. (1986). L'énonciation typographique. Communication et langages, 68. https://doi.org/10.3406/colan.1986.1762

Otlet, P. (1934). Traité de documentation. Le livre sur le livre. Les Impressions nouvelles (2015).

Perret, A., & Le Deuff, O. (2019). Documentarité et données, instrumentation d'un concept. https://hal.archives-ouvertes.fr/hal-02307039

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Schultes, E., Doorn, P., Santos, L. O. B. da S., & Dumontier, M. (2017). A design framework and exemplar metrics for FAIRness. bioRxiv, 225490. https://doi.org/10.1101/225490

Data Visualizations of Non-Human Animals and their Land

Brian Ballentine
brian.ballentine@mail.wvu.edu
West Virginia University, United States of America

Keywords: environmental humanities, animals, spatial humanities, data visualizations
Topics: English, North America, Contemporary, eco-criticism and environmental analysis, metadata standards, systems, and methods, Geography and geo-humanities, Environmental, ocean, and waterway studies

This paper presents visualizations of a large dataset containing images and metadata of nonhuman animals. Three camera traps were placed in the foothills of the Appalachian Mountains. Animals move past the camera, motion sensors trigger the taking of a digital picture and store EXIF metadata such as time and date, temperature, and moon phase. This project employs ExifTtool to manipulate image metadata, spreadsheets, Photoshop for batch resizing, and Image J and the ImagePlot macro for visualizing curated datasets. Thousands of deer, turkey, coyotes, bobcats, rabbits, raccoons, possums, and bird species were captured on camera. These individual species are tagged in datasets and their presence can be visualized using combinations of metadata. Drawing on theorists such as Iovino and Opperman, the data visualizations of animals suggest that they are not just an agential force, operating as actors amidst a multitude of others, but a narrator telling a story of the land.

The Explainability Turn: Critical Digital Humanities and Explanation

David Michael Berry
d.m.berry@sussex.ac.uk
University of Sussex, United Kingdom

Keywords: Explainability, Critical Digital Humanities, Explanation, Critical Reason
Topics: Global, Europe, English, Contemporary, digital access, privacy, and ethics analysis, meta-criticism (reflections on digital humanities and humanities computing), Media studies, Philosophy

One of the key drivers for the attention given to explainability has been a wider public unease with

the perceived bias of algorithms in everyday life, especially in the rise in automated decision processes and the calls for accountability in these systems (see Eubanks 2017, Kuang 2017, Noble 2018). Discourse and algorithms become a technique to exercise power, for example through "nudging" strategic behaviour and thereby shaping labour, both physical and mental. Through behavioural logics of control our everyday lives are subject to algorithmic management from increasingly prevalent hyper-individualised capillaries of power (Berry and Fagerjord 2017). These implications are increasingly discussed in the media and in politics, particularly in relation to a future dominated by technologies which are thought to have huge social consequences (Malabou 2019). These are important issues, but here I drill down to focus on the cognitive and explanatory issues in relation to the digital humanities. In the first half of the paper, I seek to think critically about the concept of explainability and its potential for developing a possible tactic in response to the wider toxicity generated by algorithmic governance. The aim is to offer an immanent critique of the notion of explainability. By immanent critique, I refer to an approach whereby the internal terms and concepts within a system are examined in relation to the reality of the claims they make about and the actuality of the world. In the second half of the paper I seek to show how these claims about explainability can and should be taken up by digital humanities as an explicitly interdisciplinary theoretical concern for a new research programme in the digital humanities.

I argue that the justificatory move to explainability as a panacea for computational systems is therefore an important diagnostic site for interrogating their power and ubiquity. One of the most difficult tasks facing the digital humanist today is understanding the delegation and prescription of agency in digital infrastructures. It is clear that in the context of computational systems, the first important question we need to consider is what counts as an explanation? Indeed, explanations are generally considered to be able to tell us how things work and thereby giving us the power to change our environment in order to meet our own ends. In this sense of explanation then, science is often supposed to be the best means of generating explanations (Pitt 1988: 7). So, with a stress on the importance of explanation, explainability makes it a criterion of adequacy for satisfactory use of algorithmic decision systems, and thereby legitimating their use in a multitude of settings. Thus, explainability and the underlying explanation are linked to the question of justification. I call this the "Explainability Turn." It is a genuinely interesting question as to the extent to which explainability will be able to mitigate the public anxieties manifested when confronted with opaque automated decision systems. The scale of the challenge represented by the requirement to provide an explanation seems to me to be under-appreciated, and clearing the theoretical ground for even thinking about this problem cannot be overstated.

So, what then is an explanation? Hempel and Oppenheim (1988) argue that an explanation seeks to "exhibit and to clarify in a more rigorous manner". Some of the examples they give include, a mercury thermometer which has been rapidly immersed in hot water and whole temperature reading can be explained using physical properties of the glass and mercury (Hempel and Oppenheim 1988: 10). An explanation therefore attempts to explain with reference to general laws. However, this causal mode of explanation can become inadequate in fields concerned with purposive behaviour, as with computational systems. In this case it is common for reference to purposive behaviour, such as in so-called machine behaviour, to be given in relation to "motivations" and therefore for teleological rather than causal explanation (Ruben 2016). Thus, the goals sought by the system are required in order to provide an explanation. Teleological approaches to explanation may also make us feel that we really understand a phenomenon because it is accounted for in terms of purposes, with which we are familiar from our own experience of purposive behaviour. One can, therefore, see a great temptation to use teleological explanation in relation to AI systems, particularly by creating a sense of an empathetic understanding of the "personalities of the agents." In relation to explanation, therefore, explainability can be said to aim to provide an answer to the question "why?"

The concept of explainability, and the related practices of designing and building explainable systems, also have an underlying theory of general explainability and a theory of the human mind. These two theories are rarely explicitly articulated in the literature, and I want to bring them together to interrogate how explainability cannot be a mere technical response to the contemporary problem of automated decision systems, but actually requires philosophical investigation to be properly placed within its historical and conceptual milieu. Additionally, explanation is ambiguous as it may refer to the product or to a process. It certainly seems to be the case that many discussions of explainability tend to be chiefly interested in the idea of an explanatory product. Thus, an "explanatory product" can be characterised solely in terms of the kind of information it conveys, no reference to the act of explaining being required. The question therefore becomes, what information has to be conveyed in order to have explained something? This is

something I argue that the digital humanities are well-qualified to develop as a research programme. For the user to challenge an explanation or appeal to a higher authority, if it were considered inadequate, requires expertise that presently is not readily available. I argue that digital humanities should become exemplars in the development of such systems, and by doing so contribute to a public humanities that seeks to provide knowledge and support outside of academia in relation to furthering and deepening knowledge of, and critique for, explainable systems.

Selected Bibliography

Berry, D. M. and Fagerjord, A. (2017) Digital Humanities: Knowledge and Critique in a Digital Age, Polity.

Eubanks, V. (2017) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, St Martin's Press.

Hempel and Oppenheim (1988) Studies in the Logic of Explanation, in Pitt, J.C. (ed.) Theories of Explanation, Oxford University Press.

Kuang, C. (2017) Can A.I. Be Taught to Explain Itself?, The New York Times, https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Malabou, C. (2019) Morphing Intelligence: From IQ Measurement to Artificial Brains, Columbia University Press.

Noble, S. U. (2018) Algorithms of Oppression, New York University Press.

Pitt, J.C. (1988) Theories of Explanation, Oxford University Press.

Ruben, D. H. (2016) Explaining Explanation, Routledge.

Binding Media. Contemporary Electronic Literature and Experimental Publishing

Élika Ortega
elika.ortega@colorado.edu
University of Colorado Boulder, United States of America

Keywords: electronic literature, print-digital intersections, cultural hybridity
Topics: Comparative (2 or more geographical areas), English, Contemporary, bibliographic analysis, electronic literature production and analysis, Book and print history, Media studies

There is no shortage of ways in which the print and the digital intersect within the literary realm. Binding media is a particular manifestation of these intersections. It refers to the creation and publication of literary works-fiction, essay, and poetry-using both print and digital media. Published simultaneously or sequentially in both print and digital media, this type of work reveals the tensions between media, specific modes of reading and writing, the stability of the codex and the rapid changing landscape of digital technologies, experimental writing, bookmaking, and trade publishing. Binding media works stage a contact zone and the media-cultural encounters taking place at the juncture of transnational, transcultural, and media exchanges fostered by technological shift, neocolonial dynamics, migration, neoliberal markets, and globalized communication. The examination suggests that at the turn of the 21st century, in our globalized and hyper-connected world, cultural and media encounter are best seen as a continuum.

Contextualizing Victorian Newspapers

Kaspar Beelen
kbeelen@turing.ac.uk
The Alan Turing Institute

Ruth Ahnert
Queen Mary University of London

David Beavan
The Alan Turing Institute

Mariona Coll Ardanuy
The Alan Turing Institute

Kasra Hosseini
The Alan Turing Institute

Katherine McDonough
The Alan Turing Institute

Jon Lawrence

University of Exeter

Yann Ryan
British Library

Giorgia Tolfo
British Library

Mia Ridge
British Library

Daniel van Strien
British Library

Olivia Vane
British Library

Daniel Wilson
The Alan Turing Institute

Keywords: Digital Archives, Newspaper, Bias
Topics: Europe, English, 19th Century20th Centurydigital archivingdigital libraries creation, management, and analysisHistoryLinguistics

Introduction

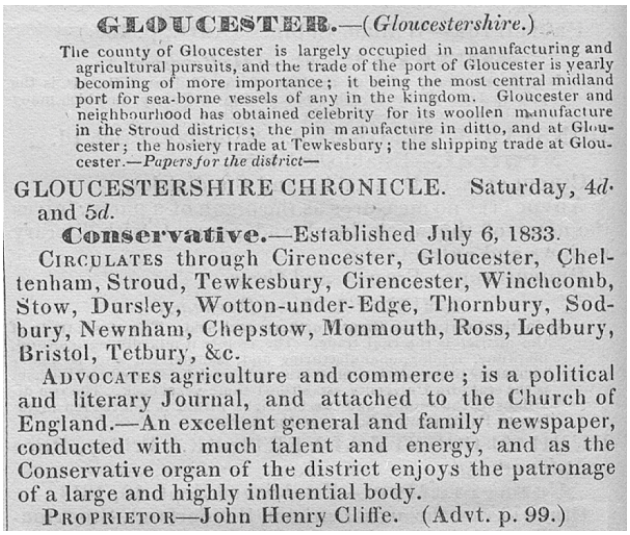
Digitised newspapers have become an increasingly prominent resource for scholars in the Digital Humanities (Bingham 2010; Smits, 2014). These digital collections provide insights into contemporaneous reporting on a wide range of events, from disruptive, historical turning points, to the banal, mundane aspects of everyday life. While rich in content, digitised newspapers by themselves often lack the essential metadata researchers need to interrogate these collections rigorously.

The British Newspaper Archive, for example, boasts that it contains more than 37 million pages, which, although amounting to a large number of words, constitutes only around 6% of the estimated total of newspaper pages. Establishing how the digital "sample" relates to the total "population" of newspapers has been difficult if not impossible. Similar challenges have been found when working with digital book collections (Francois 2013).

In this paper, we demonstrate how contextualizing newspapers from the British Newspaper Archive allows scholars to understand the contours of this corpus and to critically interrogate the digital archive for potential biases. Below, we introduce our sources, then explain how we digitized, encoded and linked them, and conclude with some preliminary results that show what type of research could emanate from this work.

The Newspaper Press Directories

To investigate the contours of the Victorian press, we set out to digitize the "Newspaper Press Directory", a contemporaneous record that catalogued periodicals circulating in Britain. Created by Charles Mitchell in 1846, the Directory was published as an annual guide for advertisers, helping them to find relevant audiences. In 1861, Mitchell's Directory was recognized by the Post Office as an authoritative list of the London and Provincial newspapers (Brake 2015; Gliserman, 1969).



GLOUCESTER.—(*Gloucestershire.*)

The county of Gloucester is largely occupied in manufacturing and agricultural pursuits, and the trade of the port of Gloucester is yearly becoming of more importance; it being the most central midland port for sea-borne vessels of any in the kingdom. Gloucester and neighbourhood has obtained celebrity for its woollen manufacture in the Stroud districts; the pin manufacture in ditto, and at Gloucester; the hosiery trade at Tewkesbury; the shipping trade at Gloucester.—*Papers for the district*—

GLOUCESTERSHIRE CHRONICLE. Saturday, 4*d*. and 5*d*.

Conservative.—Established July 6, 1833.

CIRCULATES through Cirencester, Gloucester, Cheltenham, Stroud, Tewkesbury, Cirencester, Winchcomb, Stow, Dursley, Wotton-under-Edge, Thornbury, Sodbury, Newnham, Chepstow, Monmouth, Ross, Ledbury, Bristol, Tetbury, &c.

ADVOCATES agriculture and commerce; is a political and literary Journal, and attached to the Church of England.—An excellent general and family newspaper, conducted with much talent and energy, and as the Conservative organ of the district enjoys the patronage of a large and highly influential body.

PROPRIETOR—John Henry Cliffe. (Advt. p. 99.)

Figure 1 : Example from Mitchell's Press Directory, 1857

Figure 1 shows a description of the "Gloucestershire Chronicle" taken from Mitchell's 1857 edition. The Directory systematically lists the political leaning, price and day(s) of publication, which allows us to profile newspapers, i.e. locate them geographically as well as socially (i.e. places of circulation and audience). Mitchell's attempted to provide a "dignified" account of the press, was not unproblematic. As (O'Malley, 2015) points out: the Directory didn't simply reflect the Victorian press but was also an actor that shaped the newspaper landscape.

Nonetheless, the directories provide a useful entry point to organizing the 19th century newspapers. We, therefore, attempted to contextualize the British Newspaper Archive by linking this collection to information derived from the Directory. Initially, we focussed on two counties (Lancashire and Dorset) for six different years (1847, 1857, 1860, 1867, 1877, 1888, 1898 and 1908).

We selected these two counties because this research takes place in Living with Machines, a larger project focussed on the lived experience of the industrial revolution. The counties experienced different chronologies with respect to their industrialisation: Dorset remained rural for much of the nineteenth century, while Lancashire industrialised early.

Methodologically, the biggest challenge lay in parsing and encoding these documents automatically. Mitchell's Directory adheres to a rather intricate and shifting typographical layout, with many different fonts and transitions between cursive and bold type. This, unfortunately, significantly deteriorated the OCR quality. To process these documents we trained two different models: a page segmentation model that splits the directory into separate entries (each describing one newspaper) and a "semantic" model that overlays the individual entries with more refined annotations (title, price, political leaning, but also named entities such as Location, Person and Institution). The mock example below shows the input and output of this process.

Example Input: [...] Proprietor—Henry Smith, general printer, as'the representative of Trustees. (Advertisement, p. 104.)? NEAL'S CAMBRIDGE GAZETTE. Monthly,Price U/. IMeiifr&il.—Established March, 1855. Circulates in Cambridge and its district. This is a union of the local intelligence of the district, with general news and miscellanies. Proprietor —Jonathan Neal.? CAH.TEBBI1 BY_____{Kent.)?

Example Output: <newspaper-start>[...] Proprietor—Henry Smith, general printer, as'the representative of Trustees. (Advertisement, p. 104.)? <newspaper-end>

<newspaper-start><title nid="0013824">NEAL'S CAMBRIDGE GAZETTE</title>. <publ>Monthly <publ>, Price ? <price corrected="1d">U/</price> .

<leaning corrected="neutral">IMeiifr&il</leaning> .—Established <established>March, 1855</established> . Circulates in <location lat="52.20" lon="0.12">Cambridge</location> and its district. This is a union of the local intelligence of the district, with general news and miscellanies. Proprietor—<person>Jonathan Neal</person> .? <newspaper-end>

CAH.TEBBI1 BY_____{Kent.)?

To train the models, we first manually labelled a random sample with a tool called INCEpTION (Klie et al. 2018). Structural parsing is roughly equal to page segmentation, which is reflected in the tag set (Table 1). Words that describe a newspaper (as in Figure 1) are classified as "Newspaper" (with "Start" and "Stop" indicating the boundaries). As the directories are organized geographically (by district), we tagged a new location as "District-Header" and its accompanying text as "District-Description".

Tags	Precision	Recall	F1 Score
Newspaper	1.00	0.97	0.98
Newspaper-Start	0.88	0.93	0.90
Newspaper-Start	0.85	0.90	0.88
District-Header	0.97	0.91	0.94
District-Desc.	0.99	0.97	0.98
Footer	0.90	0.66	0.76
Header	0.83	0.89	0.86
Micro Average			0.96
Macro Average			0.90

Figure 2: Table 1: Results of structural parsing

We applied Conditional Random Fields classifier to segment pages. We vectorized the target and neighbouring words using (among others) the following features: use of punctuation and capitals, part-of-speech, named-entity type, etc. We treated word-window, number of iterations and regularization terms as hyper-parameters which we selected based on a grid search. This resulted in a micro f-score of 0.96, training on a corpus of 40 pages and testing on a set of 10 pages.

Tags	Precision	Recall	F1-score
Date Establ.	1.00	0.98	0.99
Date Publ.	0.96	0.96	0.96
Location	0.96	0.94	0.95
Organisation	0.84	0.62	0.71
Person	0.79	0.89	0.84
Political Leaning	1.00	0.81	0.89
Price	1.00	0.98	0.99
Title	0.95	0.94	0.94
Micro Average			0.94
Macro Average			0.91

Figure 3: Table 2: Results semantic parsing

The semantic parser takes a segmented page as input, and establishes the "function" of each word in the document, identifying the recurrent elements of the Directory, such as title, price, and political, to which we added other basic named-entities such as Location, Person and Institution.

We followed the procedure developed by (Akbik, Blythe, and Vollgraf 2018) which applies contextual string embeddings to sequence labelling. More concretely, we used pre-trained character level Flair embeddings (both the backward and forward pass trained on a corpus of 1 billion words) to encode the tokens, which are subsequently forwarded to a Bi-LSTM. In total, we annotated 240 entries, which we split into a training (166), development (36) and test set (36).

Table 2 lists the f-scores obtained after training for 20 epochs. Currently, the tagger fails to accurately identify Organisation-type entities from the descriptions, but it does a reasonable job on other entities such as Locations and Persons. It accurately detects titles and prices but performs less well on the tokens which signal political leaning (this could be due to the many OCR errors).

After training both models we applied them to the digitised Directories. Initially, we focussed only on the section that lists provincial newspapers (given the counties of interest, we excluded London), which resulted in a set of almost 7000 newspaper profiles. Figure 4 shows the number of provincial titles over time.

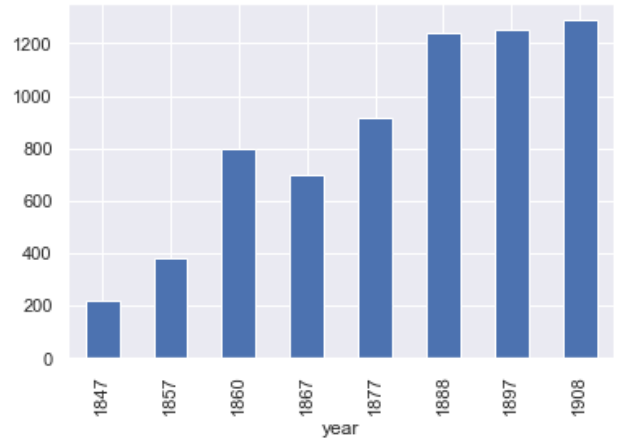


Figure 4: Number of provincial newspaper titles over time

Potential Research Outcomes

Below we discuss briefly the types of research questions we can answer by contextualizing the Victorian newspaper data. Firstly, we analysed the extent to which digitized newspapers are representative of the broader landscape with respect to their political composition.

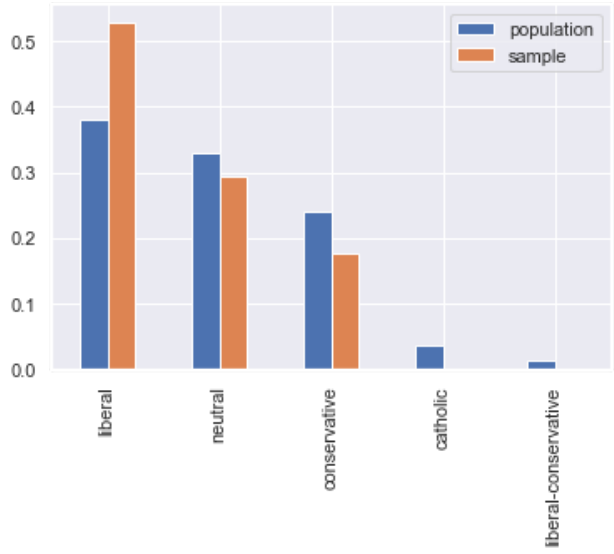


Figure 5: Distribution of political leaning in sample and population.

Figure 5 directly compares the digital sample to the newspaper population, suggesting that the distribution between them differs substantially. The Liberal press, in particular, is overrepresented compared to Neutral and Conservative newspapers (using the categorization applied by the Directory itself). Moreover, some perspectives are simply excluded, such as Catholic voices.

Future research needs to explore the impact of these biases, but preliminary experiments already hint at the possibly distorting effects on potential research outcomes and interpretations.

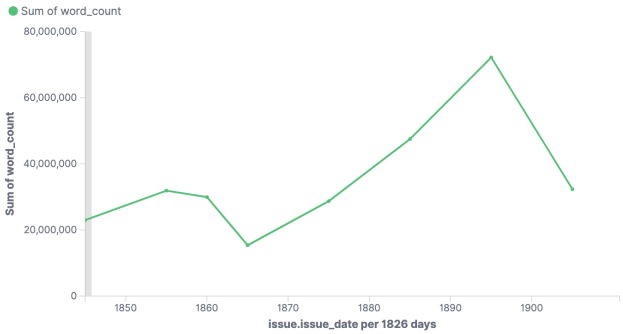


Figure 6: Number of words for articles circulating in Dorset

Figure 6 shows the word count for all the newspapers that circulated in Dorset. What this timeline fails to convey, however, is a political transition from Conservative to Liberal--which happens in the

background, and only becomes visible after integrating information from the Directory (figure 7).

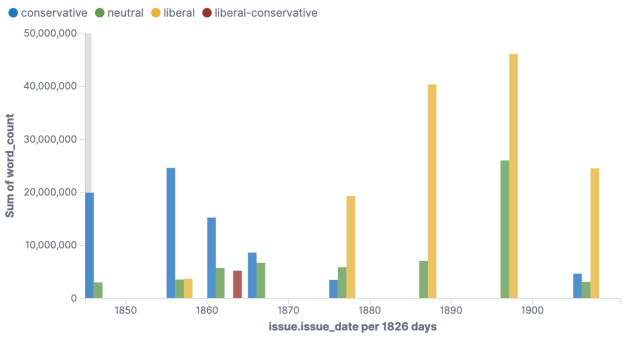


Figure 7: Number of words by political party for articles circulating in Dorset

Besides explicating the political dimension, the enriched dataset can provide a more holistic view on the geographic circulation of newspaper data as shown in Figure 8, which compares places of publication with circulation.

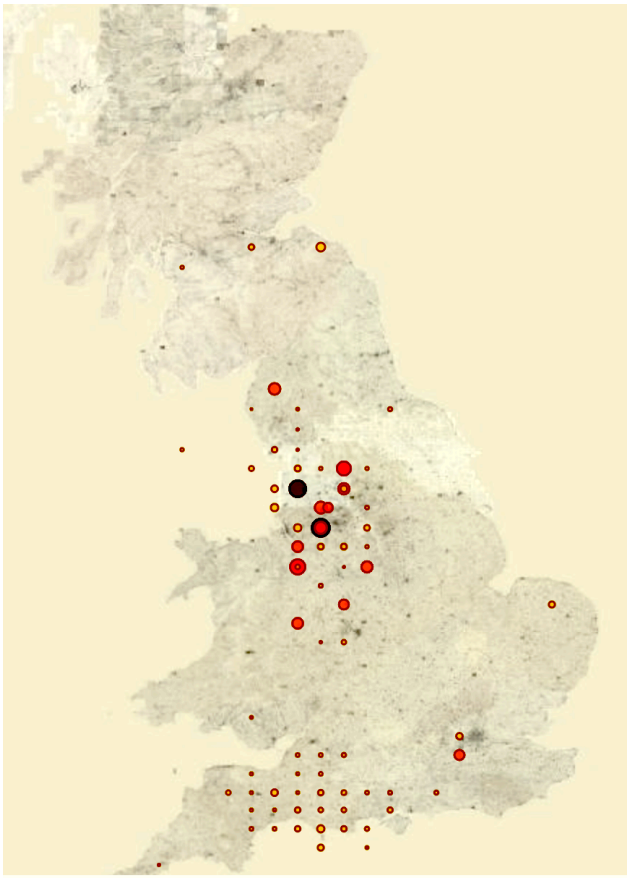


Figure 8: Number of articles by place of publication (left) and place of circulation (right) We thank the National Library of Scotland for access to their tileserver. Map images courtesy National Library of Scotland.

* Code will be available in the Living with Machines github organization: https://github.com/Living-with-Machines

References

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. 2018. 'Contextual String Embeddings for Sequence Labeling'. In Proceedings of the 27th International Conference on Computational Linguistics, 1638–49.

Bingham, Adrian. 2010. '"The Digitization of Newspaper Archives: Opportunities and Challenges for Historians"'. Twentieth Century British History 21 (2): 225–31. https://doi.org/10.1093/tcbh/hwq007.

Brake, Laurel. 2015. 'Nineteenth-Century Newspaper Press Directories: The National Gallery of the British Press'. Victorian Periodicals Review 48 (4): 569–90. https://doi.org/10.1353/vpr.2015.0055.

Francois, Pieter. 2013. 'The Sample Generator - Part 1: Origins'. British Library Digital Scholarship Blog. 2013. https://blogs.bl.uk/digital-scholarship/2013/11/the-sample-generator-part-1-origins.html.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. 'The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation'. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, 5–9. Association for Computational Linguistics. http://tubiblio.ulb.tu-darmstadt.de/106270/.

O'Malley, Tom. 'Mitchell's Newspaper Press Directory and the Late Victorian and Early Twentieth-Century Press.' Victorian Periodicals Review 48, no. 4 (2015): 591-606.

Smits, Thomas. 2014. 'TS Tools: Problems and Possibilities of Digital Newspaper and Periodical Archives', 8.

Conceptual Analysis as Thickening: Influence as an Example

Jingzhu Wei

414705062@qq.com
School of Information Management, Sun Yat-sen University, China

Allen Renear

renear@illinois.edu
School of Information Sciences, University of Illinois at Urbana Champaign, USA

Keywords: conceptual analysis, interpretive methods, digital humanities, method
Topics: Global, English, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), Logic and epistemology, Philosophy

Conceptual Analysis as Thickening: Influence as an Example

Jingzhu Wei, School of Information Management, Sun Yat-sen University Allen H Renear, School of Information Sciences, University of Illinois at Urbana-Champaign

Introduction

There has been a long conversation about method in the humanities and social sciences (Vico, Weber, Dilthey, Gadamer, and Winch, for example). Within the DH community this conversation takes a distinctive form amidst the interweaving of traditional humanistic approaches with the more empirical scientific methods associated with our new tools and research strategies[1,2,4,10]. Conceptual analysis can help support this interweaving by contributing to the forms available for the needed interpretive thickening of research in the digital humanities.

Thick description and conceptual analysis

Clifford Geertz makes a distinction, now classic in interpretative social science, between "thin" and "thick" description[6], terms first used by Gilbert Ryle, who offers this example: a *thin* description would describe a composer as producing note sequences, but a *thick* description interprets the same phenomena as "cancellings, modifyings, assemblings, reassemblings, rehearsings" - that is, the composer's thoughts and intentions situated within a particular cultural context[12].

Conceptual analysis clarifies a concept by formally identifying conditions individually necessary and collectively sufficient for its occurrence. Although central to analytic philosophy, in the last several decades conceptual analysis has also been used to complement empirical methods in the social and cultural sciences, and, more recently, in the information sciences[3,5,8,9,14,16].

For cultural phenomena the requirements identified by conceptual analysis can help provide opportunities for *thickening* descriptions. We illustrate this with the concept of (intellectual) influence.

Influence in the humanities, and in the digital humanities

Much research in the humanities explores influence - how the literary, religious, or scientific beliefs, attitudes, tastes, or feelings of a person or group of persons influenced others, often focusing on artifacts or events considered as evidence for influence, instruments of influence, or the primary agents and patients.

In the digital humanities our projects explore influence by putting data in digital form and analyzing that data with computational methods[15]. We detect stylistic similarities, transmitted corrections and errors, related themes and topics, and so on. We then advance claims or hypotheses that assert or explain th e influences we have evidence for.

Although understanding *influence* is a common research objective there has been little effort to define the concept itself. We routinely indicate what we count as evidence of influence, but rarely say, clearly and exactly, what we mean by the term - such omissions will become increasingly problematic in the future as research in the digital humanities, and in fact the humanities in general, is likely to be focusing on larger and larger quantities of thinner and thinner data.

Motivation for conceptual clarification: Causation is not Influence

We are often told that correlation is not causation - but neither is causation influence, and thinning data and methods are conducive to the latter fallacy as well. We might say a person X influenced a person Y because X's views had a causal effect on Y's views. Suppose however that a novelist's views on class lead to a successful potboiler and the proceeds are anonymously donated to a political theorist - who then has time to develop a particular analysis of class. This is causation to be sure, and a quick inference from a thin description might classify it as influence, but it is not influence - influence is causation of a *certain sort*.

The example may seem cooked up, but it makes the point. The well-known problem of "deviant causal chains"[11] is a challenge both to empirical methods, which must extract influence from causation, and to analytical efforts which attempt to identify a basis for that distinction. These are related issues. Preventing the conflation of causation and influence in a world of thin data, requires a deeper conceptual understanding of influence.

Steps towards a conceptual analysis of influence

We first adapt Grice's account of intended meaning[7]:

S states that P if and only if

S utters U intending that

- Someone, x, forms the belief that P
- x recognizes that S intends 1)
- x forms the belief that P at least partially because 2)

Grice establishes that neither 1) alone, nor 1) and 2) together are sufficient for an occurrence of some particular linguistic behavior meaning P, and although some problems remain the three clauses together to appear to be a reasonable first characterization. We therefore begin our account of influence by building on Grice:

- H is influenced by S =df
- S states that PH forms the belief that P and does so at least partially because H recognizes that S stated that P

Such a limited notion of influence may seem of little interest and we make no claims for direct applicability. A complete and robust account of influence would consider influences from/on sensibility, taste, and other affective a states, as well as capabilities, skills, propagated or transitive influence, and more; it would also include the influence of communicative objects and non-linguistic events.

Nevertheless, these analyses establishes much: the need for intention, the need for *reflexive* intention (intentions about intentions), and, in particular, for an *intention that recognition of intention* be a partial *cause* of the response. Such features can help guide the construction of thick interpretive description from thin digital data.

Deviant Causal Chains: Resolved and Restored

Another achievement of this analysis is the elimination of a class of deviant causal chain counterexamples. In the case described above the political theorist's views will fail our requirement that they be at least partially caused by the *recognition* of the novelist's views, and so will not be counted as influence.

Unfortunately another subset of cause/influence conflating counterexamples remains: suppose the novelist communicates his views on class to the political theorist, who then herself writes and publishes the potboiler (based on those views) that funds her research. The recognition condition is met, but the theorist's research has not been (intellectually) influenced by the views of the novelist.

Interpretation Abides

Eliminating deviant causal chains counterexamples has proven difficult in other analyses of fundamental social and cultural concepts, and may be impossible. But this does not undermine the practical usefulness of conceptual analysis - rather it reminds us of the limitations of formal strategies, and the fluid nature of conceptual understanding. It is significant that we rarely have any trouble with cases: once we understand the case we recognize immediately whether the causal chain is deviant or legitimizing. Conceptual analysis can help guide thick description in the digital humanities, but it is not a replacement for interpretation.

References

Ciula, Arianna, Cristina Marras, Circling Around Texts and Language: Towards "Pragmatic Modelling" in Digital Humanities. Digital Humanities Quarterly 10:3, 2016.

Clement, Tanya E., Where Is Methodology in Digital Humanities? In: Debates in the Digital Humanities, University of Minnesota, ed. Matthew K. Gold and Lauren F. Klein, 2016.

Derr, Richard L., A Conceptual Analysis of Information Need, Information Processing & Management 19:5 1983.

Domenico, Fiormonte, Valentina Martiradonna, Desmond Schmidt. Digital Encoding as a Hermeneutic and Semiotic Act: The Case of Valerio Magrelli. Digital Humanities Quarterly 4:1, 2010.

Furner, Jonathan, Conceptual Analysis: A Method for Understanding Information as Evidence, and Evidence as Information. Archival science 4:3-4, 2004.

Geertz, Clifford, Thick Description: Toward an Interpretive Theory of Culture. In: The Interpretation of Cultures, Basic Books, 1973.

Grice, H. P., Meaning. The Philosophical Review 66:3 1957.

Korman, Dan. Z., Eric Mack, Jacob Jett, Allen Renear, Defining Textual Entailment. Journal of the Association for Information Science and Technology 69, 2018.

Lee, Jin Ha, Allen H Renear, and Linda C Smith, Known?

Item Search: Variations on a Concept. Proceedings of the Association for Information Science and Technology, 43, 2006.

Liu, Alan, Where Is Cultural Criticism in the Digital Humanities? In: Debates in the Digital Humanities University of Minnesota, ed. Matthew K. Gold, 2013.

Peacocke, Christopher, Deviant Causal Chains. Midwest Studies in Philosophy 4, 1979.

Ryle, Gilbert, Thinking and Reflecting. Royal Institute of Philosophy Supplements 1968.

Shope, Robert. The Analysis of Knowing: A Decade of Research. Princeton, 1983.

Savolainen, Reijo, Emotions as Motivators for Information Seeking: A Conceptual Analysis. Library and Information Science Research 36:1, 2014.

Wei, Jingzhu, Rui Liu, An Approach to Constructing a Knowledge Graph of the Hundred Schools of Thought in Ancient China. ACM/ IEEE Joint Conference on Digital Libraries (JCDL), 2019.

Wickett, Karen. M., Allen H. Renear, Jonathan Furner, Are Collections Sets? Proceedings of the Association for Information Science and Technology 48: 2011.

Relation to conference theme: Thick description is recognized as supporting cultural understanding, especially for observers outside the studied culture; conceptual analysis adds a format that can be used for explicitly and iteratively negotiate difference and agreement.

Multilayer Network Applications in Poetry. Centralities & Automated Corpus Expansion in Computationally Assembled Anthologies

Chris Tanasescu (MARGENTO)
margento.official@gmail.com
UC Louvain, Belgium

Diana Inkpen
diana.inkpen@uottawa.ca University of Ottawa, Canada

Jean Vanderdonckt
UC Louvain, Belgium

Nicolas Burny
UC Louvain, Belgium

Vaibhav Kesarwani
University of Ottawa, Canada

Keywords: NLP, Poetry, Graph Theory, Multilayer Networks
Topics: Europe, English, North America, Contemporary, natural language processing, network analysis and graphs theory and application, Computer science, Literary studies

The paper details a few components of our international research program-Belgian and Canadian-undertaking to analyze multilayer networks and then apply these algorithms in dealing with poetry corpora. The Canadian team has developed over the past 5 years natural language processing (NLP) poetry classifiers that have been used in analyzing networked poetry corpora while the Belgian team has recently contributed algorithms for assembling multiple layers, analyzing the resulting multigraphs, and then deploying the algorithms in automatically analyzing and expanding poetry corpora.

Quantifying cultural change: literary translation in Mexican periodicals (1894-1931)

Marina Popea
marina.popea@mod-langs.ox.ac.uk
University of Oxford, United Kingdom

Keywords: Translation, cultural magazines, Mexican literature, exploratory data analysis, cultural change
Topics: South America, Europe, English, North America, 19th Century, 20th Century, database creation, management, and analysis, Literary studies, Translation studies

More than the mere transposition of words from one language to another, the practice of translation is widely seen by scholars as an instance of cultural mediation and a force that contributes to shaping culture in various ways, as translators introduce new texts and

ideas, adapting them to the target literary field.[1] Thus, translation sits at the intersection of cultures and of societies; no literature or culture exists in isolation, and the interaction between different literary systems is key to their evolution in time.[2]

Although this appears as common sense, it is a fact rarely acknowledged in literary histories, which still focus for the most part on national or, at least, linguistically homogeneous units of analysis. Taking into account the international dimension of literature and culture fosters a better understanding of the processes of cultural change, as translation is instrumental in the import of new ways of both thinking about literature and writing.[3] This kind of approach is not uncommon in Translation Studies, however, it rarely focuses on mid to long-term patterns of change. Moreover, the underlying analysis is generally of a qualitative nature, focusing on literary agents and above all on the specific elements of repertoire that are being transferred and appropriated.[4] Indeed, textual analysis is an indispensable component of such a study, and it yields fundamental insight into the particulars of the process of cultural change. It is, however, not the only tool at our disposal.

Based on a corpus of translated texts published in Mexican cultural periodicals from 1894 to 1931, this paper assesses the contribution of quantitative methods to such an analysis. It stems from my doctoral research, which outlines the evolution of the functions and uses of translation in said corpus, seeking to examine Mexican literary history from a transnational perspective. More specifically, I argue that a data-driven approach can help identify broad trends in the production of translation, which can in turn help us substantiate or challenge conclusions drawn on the basis of particular case studies, and, as such, characterise cultural change over time.

Periodical publications are fundamental for this kind of endeavour, as they are often considered as the most immediate format for intervention in public debate, including discussions around the concepts of literature and art.[5] The fact that the magazines contain competing aesthetics and ideologies make them an ideal source for the study of cultural change. This is currently being addressed in what might become a new subfield of Translation Studies, as attested by the recent special issue of *Translation and Interpreting Studies* titled "Translation and/in Periodical Publications" and the *Translation in Periodical Publications* Conference organised by the MapModern project last year.[6] Various of the pieces presented in those instances contemplate diachronic evolutions (although not necessarily focusing on literary content), and some of them rely on a data-driven approach. However, this kind of research is very incipient, and we are still facing a lack of published studies on its methodological implications and possible results.

In this context, I offer a comprehensive framework in which to approach literary translation published in periodicals, as well as a discussion of the usefulness of examining certain variables and challenges in doing so. This corresponds to the computational component of my doctoral research, which is nearing its completion. I believe that this kind of method can bring about important re-assessments and developments in literary history, especially in the Latin American context, in which these topics are hardly being addressed.

The analysis itself consists of a variety of descriptive statistics run on a database which contains metadata for each of the translations published in a corpus of nearly 20 magazines between 1894 and 1931. The variables include the issue in which each text was published, its date, title, author, translator, genre, original publication date, original language, as well as its length, starting page, and the format of its signature, in an effort to consider both the symbolic and material aspects of the publication of translations. Information about the authors and translators (such as their gender, nationality and dates of birth and death) are also recorded.

These different variables allow us to explore a considerable amount of questions, among which I have chosen to focus on the evolution of the original language of translated texts, as well as their author's nationality, and original date of publication in order to outline the cultural geography and temporality of each periodical. It is indeed telling what the cultural references are at a given moment in time, and their proximity or distance in time is an important factor when assessing the role of translation as a source of innovation. One of the most expected evolutions, to give just one example, is that of the balance between translations from French and English, which progressively shifts from the predominance of the former to the latter during the studied time period, in a process of increased hemispheric dialogue and cultural exchange. Of course, such conclusions can (and should) then be refined by considering other variables in conjunction (e.g. is this evolution shared by various literary genres or not, does it go hand in hand with a change in temporality or no, etc).

In these different areas, my paper will emphasize the key role played by both data exploration and visualisation, and the usefulness of such an approach for literary history and Translation Studies alike.

[1] Susan Bassnett, "The Translation Turn in Cultural Studies." In Constructing Cultures: Essays on Literary Translation. Susan Bassnett and André Lefevere (eds.).

Clevedon: Multilingual Matters, 1998. 123-140; Jean-Marc Gouanvic, "A Bourdieusian Theory of Translation, or the Coincidence of Practical Instances." The Translator 11.2 (2005): 147-166; Susan Bassnett and André Lefevere (eds.). Constructing Cultures: Essays on Literary Translation. Clevedon: Multilingual Matters, 1998; André Lefevere, "Why Waste our Time on Rewrites? The Trouble with Interpretation and the Role of Rewriting in an Alternative Paradigm." In Theo Hermans (ed.). The Manipulation of Literature. Studies in Literary Translation. London; Syndey: Croom Helm, 1985. 215-243.

[2] Itamar Even-Zohar, "Polysystem Studies." Special issue of Poetics Today 11:1 (1990); Pascale Casanova, "Consécration et accumulation de capital littéraire. La traduction comme échange inégal." Actes de la recherche en sciences sociales 144 (2002): 7-20; Pascale Casanova, La république mondiale des lettres. Paris: Seuil, 2008.

[3] Itamar Even-Zohar, "Polysystem Studies." Special issue of Poetics Today 11:1 (1990).

[4] Jean-Marc Gouanvic, "Pour une sociologie de la traduction: le cas de la littérature américaine traduite en France après la Seconde Guerre mondiale (1945-1960)." In Mary Snell-Hornby, Zuzana Jettmarová and Klaus Kaindl (eds.). Translation as Intercultural Communication. Amsterdam; Philadelphia: John Benjamins, 1997. 33-44; André Lefevere, "Translation Practice(s) and the Circulation of Cultural Capital: Some Aeneids in English." In Susan Bassnett and André Lefevere (eds.). Constructing Cultures: Essays on Literary Translation. Clevedon: Multilingual Matters, 1998. 41-56.

[5] Beatriz Sarlo, "Intelectuales y revistas: razones de una práctica." América: Cahiers du CRICCAL 9-10 (1992): 9-16.

[6] María Constanza Guzmán (ed.), "Translation and/in Periodical Publications." Special issue of Translation and Interpreting Studies 14:2 (2019), https://mapmodern.wordpress.com/

Null models in authorship analysis - an alternative approach to established methods in stylometry

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Keywords: stylometry, delta, null models, hypothesis testing, authorship attribution
Topics: Europe, English, 19th Century, 20th Century, attribution studies and stylometric analysis, Literary studies

In stylometry, authorship questions are currently approached in terms of two separate tasks: authorship attribution and authorship verification. Recent research in stylometry often interpreted both as classification tasks and concentrated on further improving the accuracy of these two procedures. This study proposes an alternative approach to authorship questions, treating them as hypothesis tests based on an empirical null model. The central question this approach evolves around: what does a text distance value of X actually mean for my authorship question? This essay outlines how a null model can be derived from empirical observations to answer this question. The approach allows to choose a rejection criterion for the null hypothesis that two texts have been written by different people, producing reasonable estimations for the alpha and beta error.

Legal Literacies for Text Data Mining

Stacy Reardon
sreardon@berkeley.edu
UC Berkeley, United States of America

Rachael Samberg
rsamberg@berkeley.edu
UC Berkeley, United States of America

Glen Worthey
gworthey@illinois.edu
Stanford University, United States of America

Keywords: tdm, law, copyright, fair use, ethics
Topics: English, North America, 20th Century, Contemporary, copyright, licensing, and permissions standards, systems, and processes, text mining and analysis, Law and legal studies

Our presentation, "Legal Literacies for Text Data Mining," will introduce digital humanities researchers and professionals to core skills needed to navigate law,

policy, ethics, and risk in digital humanities (DH) text and data mining (TDM) projects.

Computational text analysis and text data mining is a mainstay of DH research, but scholars often shy away from building and openly sharing diverse and representative corpora due to uncertainty around copyright and licensing restrictions for the materials they use. While misunderstanding and confusion abounds, there are few guidelines or training programs to help researchers pilot these waters. In "Legal Literacies for Text Data Mining," we will highlight five essential legal literacies (copyright, licensing, privacy, ethics and policy, and special use cases (like international collaborations)) that TDM researchers can develop to confidently compile and publish text corpora. Our presentation will build on the research we have done in creating an NEH-funded, four-day Institute (https://buildinglltdm.org/) hosted at UC Berkeley in June of 2020. The focus of our work is United States law, with an eye toward cross-boundary research.

Helping researchers and professionals build these literacies is essential for advancing knowledge in the humanities. In a recent study of the text analysis needs of humanities scholars (Green et al., 2016), participants noted that access to in-copyright texts was a "frequent obstacle" in their ability to select appropriate texts for data mining. The perception of legal obstacles does not just deter research; it biases research toward particular topics and sources of data. In response to content provider resistance, confusing license terms, and other perceived legal roadblocks, some researchers have gravitated to low-friction research questions and corpora to avoid decisions about rights-protected data. Yet their concern about working with copyrighted materials may be unfounded, as courts have found TDM methodologies that make use of copyright-protected texts to be fair uses.

When researchers artificially limit their research scope to texts that do not have access restrictions, research may be skewed to leave important questions unanswered, and the resulting TDM findings are rendered less broadly applicable. A growing body of research also demonstrates how race, gender, and other biases found in openly available corpora have contributed to and exacerbated bias in the development of artificial intelligence tools (Barocas & Selbst, 2016, Larson et al., 2016; Levendowski, 2018). At the same time, DH scholars and professionals who exercise their fair use rights in conducting TDM research may inadvertently misinterpret the scope of those rights and other legal considerations that affect how they access, store, and disseminate rights-protected works. While fair use gives clear protection to the core activity of TDM analysis, copyright and other legal regimes may nevertheless limit the inputs and outputs of that activity in important ways. Undue caution leads to missed research opportunity, but undue confidence can lead to needless risk for the researcher and her institution.

We have examined the legal contours of text data mining and identified five legal literacies for researchers and professional staff conducting or supporting TDM research. With an understanding of these literacies, researchers performing text data mining are better positioned to make legal and ethical decisions in building their corpora without fear. Currently, though, few trainings or resources integrate these legal literacies into DH TDM outreach and instruction. Moreover, our own experiences suggest that digital humanities scholars and professionals field many of the questions that arise around legal issues and TDM at the time of crisis (e.g., when university access to a database is suspended due to unlawful downloading). This places undue stress on one's ability to conduct DH TDM research and may lead institutions to unduly restrict such research via policy.

To address this gap, we created a four-day, NEH-funded institute, Building Legal Literacies for Text Data Mining (Building LLTDM), hosted by UC Berkeley during June 23-26, 2020. The goals of the Institute are to: (1) understand how law, policy, and risk management interact with digital humanities TDM projects; (2) integrate workflows for TDM research and professional support so participants can confidently pursue valuable research; (3) practice sharing their new tools and knowledge with others through exercises based on authentic consultations; (4) prototype plans for more broadly disseminating their knowledge; and (5) develop communities of practice and coordinate, where practicable, cross-institutional outreach about the TDM legal landscape. To maximize impact, all instructional materials (including sample lesson plans and exercises) will be shared publicly as a CC0 (Creative Commons Zero waiver) open educational resource (OER). This hands-on curriculum supports 32 participants (16 DH researchers and 16 professional DH support staff including librarians) and will be taught by a combination of experienced legal scholars, librarians, and researchers-all of whom are immersed in these subject literacies and workflows.

Our DH2020 presentation will elucidate these five legal literacies with reference to common DH use cases. We will also reflect lightly on the institute model and its efficacy for helping DH scholars and professionals confidently chart a course forward. Through the legal literacies we have developed, text data mining researchers are better positioned to make legal and ethical decisions in building their corpora with confidence.

The project team is composed of Rachael G. Samberg (PI, UC Berkeley), Timothy Vollmer (PM, UC Berkeley), Scott Althaus (University of Illinois), David Bamman (UC Berkeley), Brandon Butler (University of Virginia), Beth Cate (Indiana University Bloomington), Kyle K. Courtney (Harvard University), Sean Flynn (American University Washington College of Law), Maria Gould (California Digital Library), Cody Hennesy (University of Minnesota), Eleanor Dickson Koehl (University of Michigan), Thomas Padilla (University of Nevada Las Vegas), Stacy Reardon (UC Berkeley), Matthew Sag (Loyola University Chicago School of Law), Brianna L. Schofield (Authors Alliance), Megan Senseney (University of Arizona), and Glen Worthey (HathiTrust Research Center).

For more information about the Institute, see https://buildinglltdm.org.

Works Cited

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104: 671-730. Retrieved from http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf

Green, H. E., Dickson, E. F., Nay, L. R., & Zegler-Poleska, E. (2016). Scholarly needs for text analysis resources: A user assessment study for the HathiTrust Research Center. Proceedings of the Charleston Library Conference. Charleston, SC. doi: http://dx.doi.org/10.5703/1288284316464

Larson, J., Angwin, J., & Parris, T. (2016, October 19). Breaking the black box: How machines learn to be racist. ProPublica. Retrieved from https://www.propublica.org/article/breaking-the-black-box-how-machines-learn- to-be-racist

Levendowski, A. (2018). How copyright can fix artificial intelligence's implicit bias problem. Washington Law Review, 93(2): 579-630. Retrieved from http://digital.law.washington.edu/dspace- law/bitstream/handle/1773.1/1804/93WLR0579.pdf

What if we can't manage our way to better sustainability?

Katrina Fenlon
kfenlon@umd.edu
University of Maryland, College Park, United States of America

Trevor Muñoz
tmunoz@umd.edu
University of Maryland, College Park, United States of America

Keywords: sustainability, community-centered collections, digital archives
Topics: English, North America, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), sustainable procedures, systems, and methods, Library & information science

Widespread interest in increasing the sustainability of digital humanities collections has led to a growing consensus around strategies that are focused on improving managerial practices within institutions. While useful in certain contexts, managerial sustainability strategies do not convincingly address the vast majority of digital humanities collections that operate outside of institutional infrastructures. We argue that stakeholders in digital scholarship need to broaden our conception of sustainability for digital humanities collections to admit an increasingly diversified landscape of collections, and of the communities and practices responsible for building and maintaining collections. By advancing an expanded conception of sustainability, which foregrounds the communities who create collections, this paper lays the conceptual groundwork for ongoing research into the implications of alternative, collaborative, and community-centered sustainability strategies, and a fuller range of roles that cultural institutions may play to support community-determined strategies for sustaining digital humanities collections.

Reassembling Elephants: A Multi-Spatiotemporal Visualization Method for History and Humanities Data

Florian Windhager
florian.windhager@donau-uni.ac.at
Danube University Krems, Austria

Saminu Salisu
Danube University Krems, Austria

Eva Mayr
Danube University Krems, Austria

Keywords: visualization, cultural collections, multiple views, information integration
Topics: Global, English, 19th Century, 20th Century, Contemporary, cultural analytics, spatial & spatio-temporal analysis, modeling and visualization, History, Humanities computing

In the wake of digitization initiatives, we find a wide range of phenomena in historical fields of study translated into complex, event-based data. From large-scale accounts on collective developments (such as cultural histories, art-historical movements, or socio-technical translation processes) to individual biographies of actors and objects - datafication initiatives frequently utilize event-based data formats to model complex topics as constellations of historical data points. Aside from time stamps, these data points then often are defined by further attributes, which provide their geographic, categorical, or relational specifics. The availability of such complex datasets opens new opportunities for DH research and teaching, but also for the public communication of humanities topics, and for open science endeavors.

However, when engaging in the visual analysis and communication of complex historical data, scholarly or public audiences rarely get to see their multidimensional richness. Commonly, visualization tools require analysts to selectively 'cut' into the complexity of the data to highlight and project particular aspects, while neglecting other facets and data dimensions. While coordinated multiple views allow us to establish plurality of perspectives side by side (Dörk et al., 2018; Roberts et al., 2019), they come with a variety of downsides on their own, including the splitting of users' attention, and a significant amount of visual complexity (Baldonado et al., 2000; Windhager et al., 2018a). Figuratively speaking, multiple views allow us to grasp, see, and sample vital parts of the proverbial elephant, while hindering us to see the whole, dynamic organism in its particular context. If we want to overcome this state of affairs, we have to (re)connect and (re)assemble the partial impressions from multiple views for ourselves, which turns out to be a demanding cognitive task (Windhager et al., 2018a).

To provide a more integrated approach to the analysis of event-based data, we introduce the PolyCube visualization framework. As a web-based visualization system, it draws together multiple perspectives to convey a bigger picture for complex, time-oriented data, and to support synoptic exploration of the data, as well as navigation between specific perspectives for expert and casual users alike. The system provides:

- multiple spatialized (i.e. geographic and non-geographic) overview perspectives (including a map-based, a set-based, and a network-based view),
- multiple perspectives on the temporal data dimension (including space-time cube representation, juxtaposition, animation and superimposition views)
- close-up access to single events or objects on de mand, together with
- animated canvas transitions, supporting the switching between various views.

To demonstrate the analytical options of this system, we turn to the field of cultural heritage studies, and showcase the visual analysis of two cultural heritage collections (Windhager et al., 2020), including the Charles W. Cushman photography collection (Indiana University, 2007), and a corpus of influential movies, based on cinematic references (Spitz & Horvat, 2014). Objects of cultural collections are frequently connected to multiple metadata entries (such as time of origin, place of origin, class of object), or to various types of relations between objects. The PolyCube system allows for the synoptic visual analysis of these data dimension - starting from a setup of multiple coordinated space-time cubes (Fig. 1) and enabling the seamless transition to three alternative perspectives on the historical or temporal data dimension (Fig. 2).

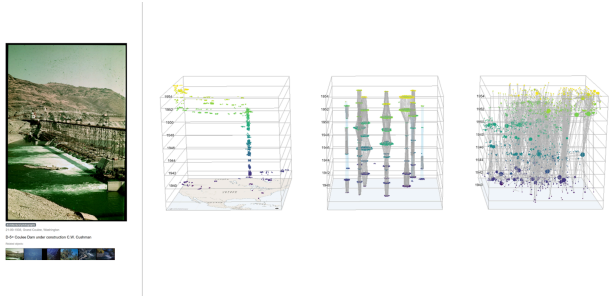


Figure 1: Three parallel views of the PolyCube system, providing a map-based, a set-based and a network-based visualization perspective (from left to right), together with a close-up view on a selected photograph.

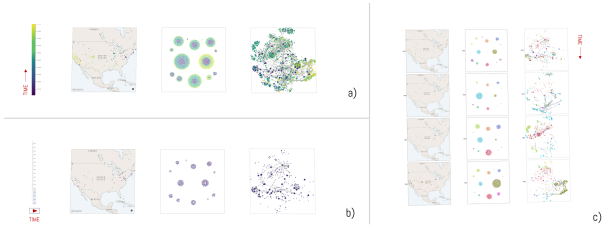


Figure 2: Three alternative views on temporal collection development, including a color-coded superimposition view (a), an animated view (b), and a juxtaposition or data comic perspective (c).

As for a general outlook, we sketch out the morphologies of a variety of further humanities topics and subject matters, which can benefit from multi-spatiotemporal representations, and for which adaptations of the PolyCube framework are under development. Transcending the point-like characteristics of cultural collection data (Fig. 3, left), actor and object biographies appear as a relevant analytical topic area. Polycubistic visualizations of traces and trajectories thus can make the movements of historical entities visually accessible in geographic, categorical and relational space-times (Mayr et al., 2020, Windhager et al., 2018b).

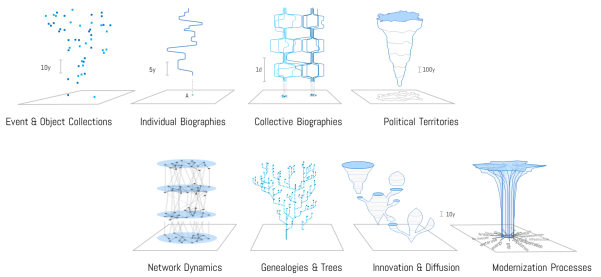


Figure 3: Outlook on a spectrum of data-driven morphologies, using space-time cube projections to make a variety of historical and conceptual phenomena visually accessible for research, teaching and public communication.

By connecting and contextualizing these basic morphological elements, data on the trajectories of political territories, on manifold actor networks, on innovation, translation and modernization processes can be remediated and rendered visible within the polycubistic framework. By creating appropriate line-ups of cubistic representations for each topic, combinatorial "meta-shapes" can appear for scholarly or public audiences, which offer bigger historical pictures for a whole spectrum of analytical or pedagogical scenarios, and around which traditional narrative accounts and argumentations can be organized. Thus we consider the suggested design to also provide one possible strategy to (re)construct and (re)assemble the shapes of elephants, hyperobjects, and miscellaneous matters of concern, which otherwise remain hidden between and behind the early stage technologies of contemporary visualization and information design (Latour, 2008).

References

Baldonado, M. Q. W., Woodruff, A., & Kuchinsky, A. (2000). Guidelines for Using Multiple Views in Information Visualization. Advanced Visual Interfaces (AVI), 110-119. https://doi.org/10.1145/345513.345271

Dörk, M., Pietsch, C., & Credico, G. (2017). One view is not enough. High-level visualizations of a large cultural collection. Information Design Journal, 23(1), 39-47. https://mariandoerk.de/papers/idj2017.pdf

Indiana University (2007). Charles W. Cushman Photograph Collection. Available at: https://webapp1.dlib.indiana.edu/cushman/index.jsp

Latour, B. (2008). A cautious Prometheus? A few steps toward a philosophy of design (with special attention to Peter Sloterdijk). International Conference of the Design History Society (pp. 2-10). http://www.bruno-latour.fr/sites/default/files/112-DESIGN-CORNWALL-GB.pdf

Mayr, E., Salisu, S., Filipov, V. A., Schreder, G., Leite, R. A., Miksch, S., & Windhager, F. (2020, forthcoming). Visualizing Biographical Trajectories by Historical Artifacts: A Case Study based on the Photography Collection of Charles W. Cushman. Biographical Data in a Digital World (BD 2019), Varna. https://osf.io/d5bse/

Roberts, J. C. et al. (2019). Multiple Views: Different meanings and collocated words. Computer Graphics Forum, 38(3), 79-93. https://doi.org/10.1111/cgf.13673

Spitz, A., & Horvát, E. Á. (2014). Measuring long-term impact based on network centrality: Unraveling cinematic citations. PloS one, 9(10), e108857. https://doi.org/10.1371/journal.pone.0108857

Windhager, F., Salisu, S., Schreder, G., & Mayr, E. (2018a). Orchestrating Overviews. A Synoptic Approach to the Visualization of Cultural Collections. Open Library of the Humanities, 4(2), 1-39. https://doi.org/10.16995/olh.276

Windhager, F., Schlögl, M., Kaiser, M., Bernád, Á. Z., Salisu, S., & Mayr, E. (2018b). Beyond One-Dimensional Portraits: A Synoptic Approach to the Visual Analysis of Biography Data. Biographical Data in a Digital World (BD 2017), 67-75. http://ceur-ws.org/Vol-2119/paper11.pdf

Windhager, F., Salisu, S., Leite, R., Filipov, V., Miksch, S., Schreder, G., & Mayr, E. (2020). Many Views Are Not Enough: Designing for Synoptic Insights in Cultural Collections. IEEE Computer Graphics & Applications. 40(3), 58-71. https://doi.org/10.1109/MCG.2020.2985368

Reading Certainty: Evidence from a Large Study on NLP and Witness Testimony

Ben Miller
b.j.miller@emory.edu
Emory University, United States of America

Keywords: natural language processing, testimony, truth and reconciliation, certainty, veridicality
Topics: Africa, Comparative (2 or more geographical areas), English, North America, 20th Century, Contemporary, cultural analytics, natural language processing, Cultural studies, Linguistics

"Reading Certainty: Evidence from a Large Study on NLP and Witness Testimony"

Ben Miller
b.j.miller@emory.edu

Witness testimony provides the first draft of history, and requires a kind of reading that connects descriptions of events from many perspectives and sources. "Reading Certainty" examines one critical step in that process, namely how a group of approximately 230 readers decided whether a statement about an event is credible and factual. That examination supports an exploration of how readers of primary evidence think about factual and counterfactual statements, and how they interpret the certainty with which a witness makes their statements. This presentation argues that readers of collections of witness testimony were more likely to agree about event descriptions when those providing the description are certain, and that the ability of readers to accept gradations of certainty were better when a witness described factual, rather than counter-factual events. These findings lead to a suggestion for how researchers in linguistics and the humanities could better model the question of speaker certainty, at least when dealing with the kind of narrative non-fiction one finds in witness testimony.

As represented by the effort that has gone in to various truth and reconciliation processes, such as occurred in Canada from 2008-2015 about treatment of First Nations in the residential schools systems, or in South Africa that from 1996 to 1998 convened hearings on abuses perpetrated during the Apartheid period, these materials serve vitally important functions for people and communities. Thus, despite their challenges, they need to be included in the types of material made more readable by computational methods. Absent any quantitative approaches, these many tens of thousands of witness statements remain mostly unread, perceived only in the aggregate form of commissions' reports.

Various challenges make the understanding of event language in real-world documents a meaningful, but difficult research problem. Often, witnesses indicate space, time, and entities referentially more so than absolutely, so their stories are resistant to techniques like named entity or temporal recognition. Additionally, these types of stories are frequently feature fragmented syntactic structure and highly referential semantics. Also, it is common that a witness either does not know where or when something specifically happened, they elide that detail, or they do not have the language with which to talk about it. In testimony provided by a first responder to the World Trade Center attacks of September 11, 2001 in World Trade Center Task Force Interview No. 9110335, an EMT says, "That's when we noticed a whole bunch of police cars responding somewhere" (Times, 2004). The resistance engendered by ambiguity hinders the critical work of associative reading, where descriptions of events by one witness are correlated to descriptions of events by additional witnesses.

While the kind of associative reading historians, linguists, and those studying narrative non-fiction could be supported by computational approaches, a critical framework to facilitate interpretation of event-related language is essential. One critical predicate for understanding the events that comprise these stories is the extent to which the speaker is certain about the statement they put forward. Any question about the computational reading of event language is conditioned on whether or not it might be something an algorithm can be trained to identify, in addition to whether that perspective is meaningful relative to understanding the witness, their statement, and the event to which they offer testimony. This measure of speaker certainty is also known as veridicality. A second critical predicate in relation to this material's role as an anchor of collective memory is the speaker's statement's facticity. Combined, these measures provide a first step in ascertaining whether the description of an event in one witness testimony can be legitimately connected to an event description from another testimony.

To explore this idea, a large study was conducted that asked; how are certainty and uncertainty indicated in the language of witness statements, and how do readers interpret those statements. This research builds on the work of (Hyland, 2005); (De Marneffe et al., 2012); (Saurí and Pustejovsky, 2009, 2012);

(Lee et al., 2015); and (Stanovsky et al., 2017) in the area of using NLP to quantify a speaker's certainty about their statements. Following the above examples, a questionnaire was created to gather judgments of veridicality from Amazon Mechanical Turk users. mTurk is a cloud-labor platform that connects workers with information processing tasks. For this task, first, sentences were extracted from different corpora of interviews from different contexts: the South African Truth and Reconciliation Commission, the Cambodian Khmer Rouge Tribunal, interviews with survivors of the Holocaust, statements from survivors of the Rwandan genocide, and interviews with survivors of ethnic cleansing in the former Yugoslavia. One goal of this study was to use real-world data, rather than simulated data. While it can be argued that simulated data would allow for a stronger statement to be made about the quantitative findings and sources of variation, it wouldn't reflect how witnesses use language, or how readers grapple with the complex problem of understanding witness testimony. A random sample of sentences containing about 800 events was taken from each of the seven corpora leading to a total of approximately 4,000 events. A total of 227 unique raters participated in the study providing a total of 27,800 individual event ratings. Those ratings were categorical, based upon the existing best model for assessing certainty according to the literature referenced above. However, a majority of raters were only likely to agree on the category when the event was considered to be Certain Fact, or Probable Fact. Table 1 below indicates the initial categories and when at least half of the raters of an item agreed.

Certain +	1083
Probable +	370
Possible +	21
Certain -	18
Probable -	5
Possible -	0
Certain by under-specified	0
Uncertain	0
Error	0

Table 1: Simple Majority Interrater Agreement

These findings suggest that NLP and computational social sciences should process assessments of veridicality with more allowances for counterfactual or uncertain statements, and more gradations of certainty and positive evidence. To better capture the relative certainty and facticity implied by those ratings, an alternative to simple majority agreement was implemented. Instead, the means of all valid ratings were calculated, then plotted where the x-axis denoted a continuum from counterfact, or event-negation, to fact, or event, and the y-axis denoted certainty. Those plotted results, shown in Figure 1, were clustered using a number of clusters determined by the elbow method, wherein the number of clusters is increased until an elbow appears in the graph. Based on these clusters, as shown in Table 2, I propose a new nine-element schema for the evaluation of veridicality. This new schema better reflects that better reflects how readers interpreted real-world testimony and the evidence from this study, and potentially better describes how readers process information about certainty and events in witness testimony.

Cluster	Within SS	Variance	SD	Nomenclature	Certainty	Counter/Fact	No. of Items
6	4596.2	71.8	8.5	Certain Fact	27.6	26.6	483
3	7625.9	119.2	10.9	Probable Fact	22.9	20.2	581
5	5062.1	79.1	8.9	Likely Fact	16.6	15.6	451
7	4798.2	75.0	8.7	Possible Fact	10.9	10.0	336
8	3843.4	60.1	7.7	Certain Unknown	21.9	5.1	117
2	4106.8	64.2	8.0	NA / Other	2.6	4.6	173
1	3846.9	60.1	7.8	Uncertain Unknown	-9.7	-0.9	64
9	5068.0	79.2	8.9	Possible Counterfact	9.5	-7.9	113
4	8555.1	133.7	11.6	Certain Counterfact	23.4	-16.5	172

Table 2: K-means Clustering of Survey Results

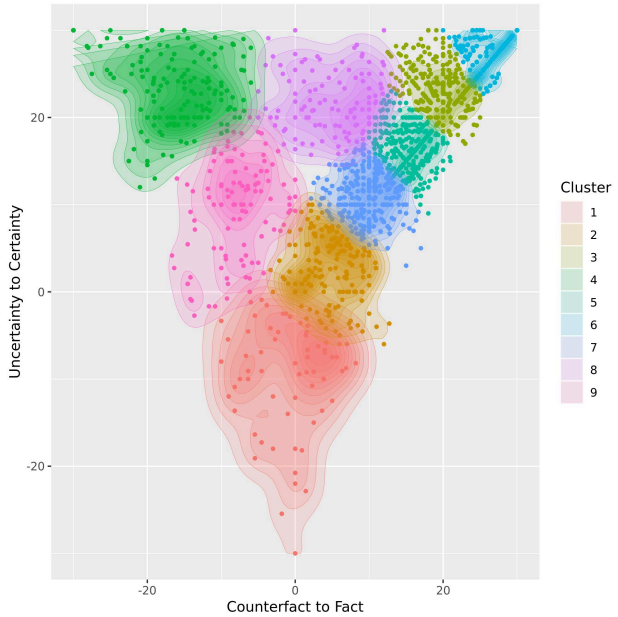


Figure 1: Clusters of Certainty to Uncertainty and Factual to Counterfactual Statements

With this theoretically and empirically developed instrument, linguists, historians, and readers of narrative non-fiction such as witness statements, may be better able to move forward with methods for computational approaches to associative reading, and better able to reengage with the primary evidence of our societies truth and reconciliation processes.

Works Cited

De Marneffe, Marie-Catherine, Manning, Christopher D, and Potts, Christopher. 2012. Did it happen? The pragmatic complexity of veridicality assessment. Computational linguistics, 38(2), 301-333.

Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. Discourse studies, 7(2), 173-192.

Lee, Kenton, Artzi, Yoav, Choi, Yejin, and Zettlemoyer, Luke. 2015. Event detection and factuality assessment with non-expert supervision. Pages 1643-1648 of: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.

Saurí, Roser, and Pustejovsky, James. 2009. FactBank: a corpus annotated with event factuality. Language resources and evaluation, 43(3), 227.

Saurí, Roser, and Pustejovsky, James. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. Computational Linguistics, 38(2), 261-299.

Stanovsky, Gabriel, Eckle-Kohler, Judith, Puzikov, Yevgeniy, Dagan, Ido, and Gurevych, Iryna. 2017. Integrating deep linguistic features in factuality prediction over uni_ed datasets. Pages 352-357 of: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Times, The New York. 2004. The Sept 11 Records. The New York Times, Aug.

Novel Worldbuilding: Science Fiction

Nichole Misako Nomura

nnomura@stanford.edu
Stanford University

Mark Algee-Hewitt

mark.algee-hewitt@stanford.edu
Stanford University

Keywords: Worldbuilding, Microgenres, Science Fiction, NLP, Improbability
Topics: English, North America, 20th Century, Contemporary, text mining and analysis, Literary studies

Novel World-Building: Science Fiction

Introduction

This project explores two narrative techniques that allow authors of Science Fiction (SF) to create and communicate invented worlds.

Explicit Worldbuilding: Microgeneric worldbuilding

Explicit worldbuilding--moments that appear to merely convey worldbuilding information--is simultaneously widely acknowledged as a unifying feature of the genre[1] yet derided as "witless, even infantile."[2] Questions of when and how much explicit worldbuilding happens, then, might help us answer questions about prestige and more fundamental questions about the form of SF.

After initial feature-finding forays into explicit worldbuilding, we turned to the Lit Lab's Microgenres project,[3] hoping to replicate their work on a specifically SF corpus. The Microgenres project seeks to identify extra-disciplinary discourses within narrative using a non-lexical approach. We hypothesized that moments of explicit worldbuilding might resemble these disciplinary discourses.

Corpus

To test this hypothesis, we assembled a test corpus of 17 SF novels and 26 science texts written by Isaac Asimov, sampled to comparable size.

Methods

With the Microgenres feature set (including frequency of Penn Treebank POS tags, average sentence length, average number of clauses per sentence, and numbers of named entity persons), we created a classification model using linear discriminant analysis. By training a model on 20- and 50-sentence subsections of our corpus of science writing and then classifying similarly-sized passages of SF, we can use the posterior probabilities of the classification results to identify the mixture of science writing in each part of each SF novel.

Analysis

This section of the project seeks to answer two related questions. First, are there significant stylistic differences in Asimov's science and science fiction? And, second, if those stylistic differences exist, can we identify moments where science style appears in SF?

To answer the first question: yes. Our classification model had a success rate of 95% at the 50-sentence level and 91% at the 20-sentence level (*Figure 1*).

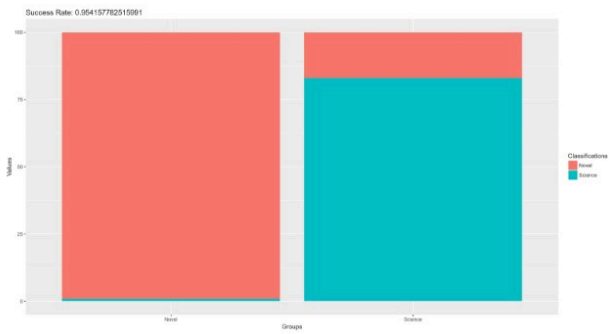


Figure 1: Classification success rate, 50-sentence slices

To answer the second question, we classified slices in each of the novels. For example, in the resulting graph of *Second Foundation* (*Figure 2*), each bar represents a 50-sentence slice of the text and the distribution of colors indicates the posterior probabilities of each discourse in each slice as assigned by the model. Slice 36, with the most "science," contains an essay on the history of the Foundation.

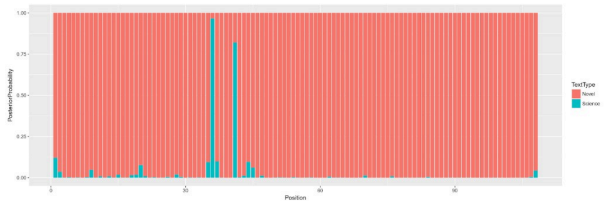


Figure 2: Classification of 50-sentence slices of Second Foundation

(Im)probabilities and Worldbuilding

In implicit worldbuilding, authors juxtapose tokens that are familiar to readers within contexts in which their co-occurrence is unexpected, such as"the door dilated."[4] While both "door" and "dilated" are familiar to readers, their unexpected co-occurrence signals a new world.

Corpus

To explore the role of syntagmatic pairs in implicit worldbuilding we expanded our corpus to include 246 SF novels published between 1905 and 2017, which we compared to a combined corpus of 146 novels from the same period tagged as "realism" and a corpus of 311,580 journal articles from *Scientific American* and the *Journal of the British Medical Association (JBMA)*. [5]

Methods

We tested normalized pointwise mutual information (NPMI) as a means to identify such bigrams as above, but the sensitivity of PMI to low frequency words (rather than low frequency word pairings) made it unable to detect bigrams of the kind we sought. Our interest, in this particular project, is in improbable or novel combinations of otherwise normal frequency words--and the signal of those improbably-combined words are drowned out by the noise created by pairings of low-frequency words.

In order to identify implicit worldbuilding, we introduce a new metric, improbability, as a way of measuring the significance of word pairs whose constituent elements occur in reference corpora, but whose combination is relatively unique to our target corpus.

Subtraction of the probability of words x and y following each other in our reference corpus from our target corpus gave too much significance to instances in which rare tokens in the non-SF corpus skewed the probability of their co-occurrence. Accordingly, we scaled the resulting metric using the zeta measure of significance for the terms.[6] This adjusted our metric to account for the relative significance of the terms in our corpora as described in Figure 3.[7]

$$\left(p(x_i, y_i)(log_2(\zeta x_i) + log_2(\zeta y_i))\right) - \left(p(x_j y_j)\left(log_2(\zeta x_j) + log_2(\zeta y_j)\right)\right)$$

Results

The 15 most improbable word combinations indicative of SF when compared to the realism and JStor corpus are shown in *Table 1*. Our metric captures both syntagmatic worldbuilding ("haziness mars") and a semiotics of science which places familiar objects in unfamiliar narrative contexts ("bunsen burner").

Word1	Word2	Improbability
bunsen	burner	6.39231742
tollgates	ish	5.4918531
circularity	nessus	5.45943162
coxcomb	lyra	5.42626475
haziness	mars	5.04439412
bhakti	yoga	5.03157917
windbag	mallow	4.64385619
dotard	toran	4.52356196
hellenic	corporal	4.52356196
whippings	toran	4.52356196
backfire	ish	4.4918531
tau	ceti	4.46080963
clouted	lyra	4.42626475
elliptically	polarized	4.32192809
billions	of	4.31440735

Table 1: Most improbable bigram sequences in SF novels

We are also able to calculate the improbability score of all bigrams in a given segment of text. This allows us to locate segments of novels with a large number of improbable word combinations, not only revealing passages of specific novels in which implicit worldbuilding occurs, but laying the groundwork for uncovering larger patterns across novels.

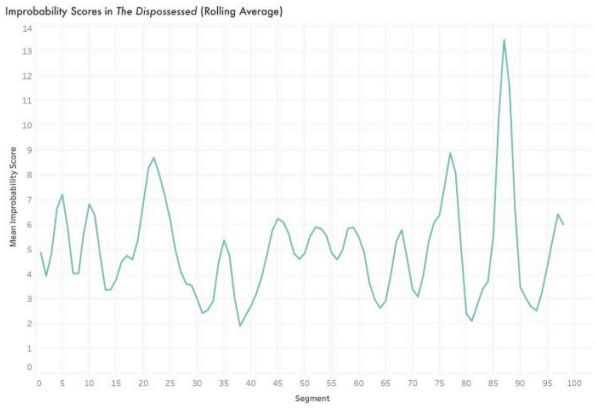


Figure 4: Improbability scores across The Dispossessed

For example, *Figure 4* shows a smoothed plot of the improbability scores across Ursula K. Le Guin's *The Dispossessed*. We divided the novel into 100 overlapping segments, calculated the improbability score in each, and averaged the results across windows of 3 adjacent segments, revealing two peaks. Our metric captured these important moments of worldbuilding, both of which depict the narrator struggling to make sense of his situation.

Conclusions

The two methods we employ, microgenres and improbility, have proven successful at identifying key moments of worldbuilding in SF. More importantly, the two metrics correspond to explicit and implicit worldbuilding, creating the opportunity to study not just individual examples of how these two strategies are employed by authors at the level of the text, but also patterns that differentiate SF from other literary genres.

Endnotes:

[1]Russ, "Towards an Aesthetics of Science Fiction." (1975)

[2] Angenot, "The Absent Paradigm: An Introduction t o the Semiotics of Science Fiction." (1979)

[3]Algee-Hewitt, Mark, Michaela Bronstein, Abigail Droge, Erik Fredner, Eva Grant, Ryan Heuser, Xander Manshel, Nichole Nomura, J.D. Porter, and Kelsey Reardon. (2019). "Microgenres: A Computational Model of Disciplinarity and the Novel." Presentation, Alliance o f Digital Humanities Organizations Conference, Utrecht.

[4] Ellison, Harlan. "A Voice from the Styx," (1968)

[5]JStor Data For Research (https://www.jstor.org/dfr/). Accessed: October, 2018

[6]Schöch, Christof, Daniel Schlör, Albin Zehe, Henning Gebhard, Martin Becker, Andreas Hotho. "Burrows' Zeta : Exploring and Evaluating Variants and Parameters" Digital Humanities 2018 (July, 2018)

Evert, Stefan, Fotis Jannidis, Thomas Proisl, Steffen Pielström, Thorsten Vitt, Christof Schöch, and Isabella Reger. (2017) "Understanding and Explaining Distance Measures for Authorship Attribution." Digital Scholarship in the Humanities (https://academic.oup.com/dsh/article-pdf/32/suppl_2/ii4/21298943/fqx023.pdf)

Burrows, John. (2006). "All the way through: Testing for Authorship in Different Frequency Strata." Literary and Linguistic Computing 22(1): 27-47.

[7] We follow Schöch et al., 2018 in using a log transformation of our zeta scores to increase their significance.

Teaching Digital Humanities on Raspberry Pis. A Minimal Computing Approach to Digital Pedagogy.

Wout Dillen
wout.dillen@uantwerpen.be
University of Antwerp, Belgium

Joshua Schäuble
Joshua.Schaeuble@uantwerpen.be
University of Antwerp, Belgium

Keywords: digital pedagogy, minimal computing, raspberry pi, iiif, digital scholarly editing
Topics: Global, Europe, English, Contemporary, curricular and pedagogical development and analysis, scholarly editing and editions development, analysis, and methods, Education/ pedagogy

As this year's conference theme suggests, Digital Humanities finds itself on the intersection between many different "disciplines, methodologies, and pedagogies" (Estill and Guiliano 2019). But while this intrinsic interdisciplinarity (see e.g. Klein 2015) can be regarded as one of the research field's chief assets, it quickly becomes a considerable hurdle in the classroom. As Ryan Cordell suggests, we tend to overestimate the technical skills of our 'tech-savvy,' 'digital native' students - and perhaps even the general affinity they have towards digital tools and methodologies (Cordell 2016, 465). Without a formalized technical or computational background, the skills our students bring to the table rely heavily on their individual interests, experiences, and auto-didacticism. This situation culminates in a heterogenous classroom that makes it difficult to assess where the gaps in each of the students' knowledge lie, and how they can be bridged. As such, the teacher's first task becomes to decide which basic skills the students will need for their digital research. As a result, technical DH courses often focus on teaching students specific skills that can be used in specific contexts to achieve specific research results, such as: How to use TEI-XML to encode a text-bearing object; How to use python to query a corpus of digital texts; How to use R to perform a stylometric analysis; How to use GIS frameworks to map data layers onto geographical zones; etc.

As Mahony and Pierazzo argued in 2012, however, more than teaching such purely technical skills there is "a need for teaching methodological approaches" (215). Rather than teaching students to use new technologies, we need to teach them how those technologies work: how their methodological constraints affect our data; how they determine what we can and cannot do with our data; and how seemingly simple data processing steps imply "interpretative decisions that change the nature of the data and how it can be validly interpreted" (Melgar et al. 2019; see also van Zundert and Dekker 2017). Still the question remains: "What are the right basic requirements" for our students? (Scagliola et al. 2019) - Where do we start?

In this paper, we propose a 'minimal digital pedagogy' that applyies the principles of Minimal Computing in the classroom. As a working group of ADHO's GO:DH Special Interest Group, Minimal Computing sets out to rethink DH work for areas in the world where factors such as high-end hardware, software, network capacity, power, etc. are not a given, by performing our computational research "under some set of significant constraints."[1] It is precisely such constraints, we believe, that allow us to get to the fundamentals of the technologies we are using - and to give our students a critical understanding of the tools they are using by building and configuring them from the ground up.

This is the setup we have adopted for the Antwerp Summer School in DH[2] - an intensive one-week summer school that aims to teach students the basics of developing and hosting their own digital archives and editions. In the last two years, we have developed a tutorial for deploying a IIIF-compliant image server on a local network of Raspberry Pis (RPi). At the next iteration of the summer school, we want to take this setup one step further by teaching students how develop a full-fledged Digital Scholarly Edition based on those reproductions.[3]

Much to our continued surprise, experience tells us that all of this is possible on a RPi. These affordable light-weight minicomputers the size of a credit card were originally designed to teach digital novices (especially children) computing and digital making skills, but have recently been successfully used in higher education classrooms as well (e.g. Zimmermann, Caballero Jr, and Brown 2017; Pasolini, Bazzi, and Zabini 2017). In our specific setup, students used RPis to up a Local Area Network where they installed, configured, and controlled web-based tools to host, share, and reuse digital reproductions of manuscripts with each other. This collaborative act of building a private network

offered them a practical way to learn more about and play with fundamental computing concepts. In particular, through a series of tutorial instructions, integrated exercises, and group discussions, they were exposed to issues in physical computational and networking infrastructures; to working with unfamiliar operating systems; to operating a computer without an interface (command line); to communicating with and controlling other computers (SSH); to a number of data formats and markup languages (HTML, XML, CSS, JSON, Pyramid TIFF, etc.); and - crucially - they learned to approach important technological standards and complex digital frameworks (such as IIIF) from a developer's perspective rather than from a user's.

But alongside the advantages of learning-by-building (Clement 2012; Ramsay and Rockwell 2012) and its applicability to a range of DH technologies, adopting a minimal computing approach to DH pedagogy has other benefits as well. To start, the setup is cost efficient: it requires no proprietary software, and can be run quasi-independently of the host institution's technical infrastructure. A such, the minimal set of hardware for this course requires a onetime investment of around $1,500 - or less than a single up-to-date desktop computer. What is more, if the host institution provides an internet connection without AP isolation and participants bring their own laptops, the entire hardware setup for a course with up to 30 participants may even fit into a standard-sized carry-on trolley. And for students without personal computers, minor adaptations to the tutorial could be considered to provide a similar learning experience. Finally, since our current setup only uses the internet to update and download free Linux packages (which might be provided via a local repository within the RPi network), it may even be possible to teach this course without an internet connection. As such, this portable and independent setup would allow us to teach DH in areas with limited access to digital infrastructures - a setup that we hope to try out at a collaborating university in the Philippines.

Notes

[1] See: https://go-dh.github.io/mincomp/about/.

[2] See: https://www.uantwerpen.be/en/summer-schools/digital-humanities/.

[3] Preliminary testing assured us that our setup can handle eXist-db 5.0.

Bibliography

Cordell, Ryan. 2016. "How Not to Teach Digital Humanities." In Debates in the Digital Humanities 2016, edited by Matthew K. Gold and Lauren F. Klein. University of Minnesota Press, pp. 459-474.

Clement, Tanya. 2012. "Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles, and Habits of Mind." In Digital Humanities Pedagogy. Practices, Principles and Politics, edited by Brett D. Hirsch. Open Book Publishers, pp. 365-388.

Estill, Laura, and Jennifer Guiliano. 2019. "CFP: English - DH2010." https://dh2020.adho.org/cfps/cfp-english/.

Klein, Julie Thompson. 2015. Interdisciplining Digital Humanities. Boundary Work in an Emerging Field. University of Michigan Press, Digitalculturebooks.

Mahony, Simon, and Elena Pierazzo. 2012. "Teaching Skills or Teaching Methodology?" In Digital Humanities Pedagogy. Practices, Principles and Politics, edited by Brett D. Hirsch. Open Book Publishers, pp. 215-225

Melgar Estrada, Liliana, Mari Wigham, and Marijn Koolen. 2019. "Programming Humanists - What is the Role of Coding Literacy in DH and Why does it Matter? Panel presented at DH Benelux 2019 (13 September 2019, Liège). http://2019.dhbenelux.org/wp-content/uploads/sites/13/2019/08/DH_Benelux_2019_paper_25.pdf [accessed 22 October 2019].

Pasolini, Gianni, Alessandro Bazzi, and Flavio Zabini. 2017. "IEEE Signal Processing Magazine." A Raspberry Pi-Based Platform for Signal Processing Education, no. 34:4 (July): 151-158.

Ramsay, Stephen, and Geoffrey Rockwell. 2012. "Developing Things: Notes toward an Epistemology of Building in the Digital Humanities." In Digital Humanities Pedagogy. Practices, Principles and Politics, edited by Brett D. Hirsch. Open Book Publishers, pp. 75-84.

Stefania Scagliola, Susan Aasman and Gerben Zaagsma. 2019. "From Idea to Implementation: Manifesto for integrating Digital Humanities into the academic curriculum - phase 2." Workshop presented at DH Benelux 2019 (10 September 2019, Liège). http://2019.dhbenelux.org/wp-content/uploads/sites/13/2019/07/DH_Benelux_2019_paper_77.pdf [accessed 22 October 2019].

van Zundert, Joris J., and Ronald Haentjens Dekker. 2017. "Code, scholarship, and criticism: When is code scholarship and when is it not?" Digital Scholarship in the Humanities, 32(1): 121-133, https://doi.org/10.1093/llc/fqx006

Zimmermann, Margaret, Jesus Caballero Jr, and Richard Brown. 2017. "Educational Videos to Teach Parallel Computing on the Raspberry Pi." In The Midwest Instruction and Computing Symposium 2017.

Recognizing Indigenous Sovereignty and Imagined Futures in the Context of Linked Data

Stacy Allison-Cassin
sacassin@yorku.ca
York University, Canada

Dean Seeman
dseeman@uvic.ca
University of Victoria, Canada

Keywords: Linked Open Data
Topics: English, North America, Contemporary, data modeling, linked (open) data, Library & information science

This paper outlines work to on the creation of a LOD store and data model related to Indigenous vocabularies using the Wikibase software.

"Work In Progress": Public Humanities, Digital Humanities, and Forms of Labor, Collaboration, and Audience

Jim McGrath
james_mcgrath@brown.edu
Brown University, United States of America

Keywords: public humanities, labor, project development, digital pedagogy
Topics: English, North America, Contemporary, digital ecologies and digital communities creation management and analysis, public humanities collaborations and methods, Cultural studies, Galleries and museum studies

Drawing on experiences as a postdoctoral fellow in Digital Public Humanities and the recent and ongoing turn toward public humanities in digital humanities, this long paper reflects on the pace of digital humanities projects and the privileging of "work in progress" and their impact on nonacademic audiences and collaborators as well as practitioners in precarious or contingent academic positions.

Gamergate: Predicting the Present

Geoffrey Rockwell
grockwel@ualberta.ca
University of Alberta

Robert Budac
University of Alberta

Bettina Berendt
KU Leuven

Florence Chee
Loyola University Chicago

Todd Suomela
Bucknell University

Keywords: Gamergate, Game Studies, Social Media
Topics: English, North America, Contemporary, social media analysis and methods, text mining and analysis, Cultural studies, Games studies

In August of 2014 a disgruntled ex-boyfriend of a female game designer posted a vicious diatribe on the net accusing her and the gaming media of corruption. This launched a culture war in the gaming community that played out on Twitter and elsewhere under the rubric #Gamergate. This paper will present key results from the study of the over 11 million tweets gathered.

Scrollership: a new name for some new ways of writing

Benjamin M Schmidt

bs145@nyu.edu
New York University, United States of America

Keywords: Scrollership, Scrollytelling, Digital narrative, digital publication.
Topics: English, North America, Contemporary, data publishing projects, systems, and methods, digital publishing projects, systems, and methods, Geography and geo-humanities, History

While digital humanities projects have developed a vibrantly broad vocabulary of new forms of presentation for data, a relatively impoverished default idea of scholarship as books and articles maintains a strong hold even here. I argue that this is a bad thing; and that many of the pieces are in place for a form of a long-form scholarly communication that better takes advantage of the possibilities for interaction using Javascript and data storage on web servers.

In homage to the slightly inane but undeniably catchy term "scrollytelling" that has come to capture much digital media practice, I dub a humanities counterpart "scrollership," and identify the ways that humanistic interactive argumentation can build on models created for journalism, the sciences, and the social science. While reproducible research and interactive publications are often thought of as two different things, in the Digital Humanities, they can better be conceived of as paired goals.

Morgan Papers: Exploring the Correspondence of California's First Female Architect

Zach Vowell
zvowell@calpoly.edu
California Polytechnic State University, United States of America

Ethan Kusters
California Polytechnic State University, United States of America

Luca Soares
California Polytechnic State University, United States of America

Samuel Verkruyse
California Polytechnic State University, United States of America

Joey Wilson
California Polytechnic State University, United States of America

Foaad Khosmood
California Polytechnic State University, United States of America

Keywords: Julia Morgan
Topics: English, North America, 20th Century, text mining and analysis, user experience design and analysis, Computer science, Library & information science

Descriptive metadata and full-text transcripts have long been valued for their roles in powering search engines and faceted browsing. But as the morganpapers.org web application demonstrates, such textual data (both structured and unstructured) can be leveraged to build a variety of tools which provide deeper and broader insight than simple searching and browsing.

Cal Poly recently completed digitization of a unique body of correspondence between architect Julia Morgan and William Randolph Hearst, carried out during the construction of what is now known as Hearst Castle. The structure is a masterpiece and the crown jewel of Morgan's illustrious career throughout California, where she worked as the state's first female licensed architect. The collection consists of over 2,500 letters, telegrams, notes, and other documents, spanning the years 1919-1941. As each piece of correspondence was digitized, it was ingested in the library's archival repository along with its MODS-based metadata, and full-text transcripts.

Speech Rhythm and Syntax in Poetry and Prose

Thomas Nikolaus Haider
thomas.haider@ae.mpg.de
Max Planck Institute for Empirical Aesthetics, Frankfur University of Stuttgart

Keywords: Rhythm, Meter, Speech Rhythm, Syntax

Topics: Europe, English, 18th Century, 19th Century, 20th Century, natural language processing, text mining and analysis, Linguistics, Literary studies

Abstract.

We perform several experiments regarding the interaction of local syntax with meter and rhythm in poetry and prose. To that end, we annotate 3600 lines of German poetry for syllable stress, foot boundaries, caesuras, and the main accents in a line, to then train Conditional Random Fields to predict these features and also part-of speech. With these models, we first determine a stress accent ratio, i.e., the likelihood that a particular part-of-speech is stressed or unstressed in a larger poetry corpus, which then allows us to establish a stress hierarchy (nouns are usually stressed, but determiners seldom so). Second, we determine the context dependence of these stress ratios by investigating when and how strongly stress ambiguous words (like adverbs) change their stress ratio based on their local context. Third, we classify poetry against prose, both on a syntactic and a rhythmic level, to then interpret feature weights. Finally, we look at the interaction of a simple form of enjambement with part-of-speech and verse measures.

1 Introduction

The verbal rendering of thought requires the choice of appropriate lexical items and their ordering according to the rules of syntax. Syntax, however, does not fully determine word order: speakers and writers can often choose among possible syntactic constructions when formulating their message. Semantic, pragmatic, as well as phonological constraints are known to affect wording. In spontaneous language production, semantic constraints presumably control sentence structure more immediately and to a stronger degree than phonological constraints. This follows from the logical directionality of language production, in which the semantic content of the message governs lexical choice and the assignment of syntactic function; phonology and rhythm can exert their role and endow the structure with sound only once a syntactic scaffold has been constructed (Levelt, 1993).

In contrast, the diction of poetry can be strongly governed by prosodic regularities that can be used not only for aesthetic effect by endowing a musical quality (Menninghaus et al., 2017; Menninghaus et al., 2018), but also through deliberate deviation from a regular form, consequently highlighting the relevant utterance (Attridge, 2014; Blohm and Menninghaus, 2018). In order to enforce a particular prosodic form may thus affect the choice of syntactic constructions and the order of constituents within a sentence, attesting the influence of prosody on syntax in lyric poetry.

To analyze the interaction of syntax with meter and rhythm at scale, we make use of the German poetry corpus in version 4 (Haider and Eger, 2019), containing the poetry of Textgrid (textgrid.de) and the German text archive (DTA: deutschestextarchiv.de). The corpus contains 59k poems with over 1.6M lines. It is available at github.com/tnhaider/DLK. See Table 1 for a size overview.

We train Conditional Random Fields (CRF) with the sklearn crf-suite to automatically annotate the whole German Poetry corpus (1.6 million lines) for part-of-speech (POS), binary meter (BM), and natural speech rhythm with three levels of syllable prominence (TR). To that end, two students of linguistics / literature manually annotated 3600 lines of school canon poetry for binary meter (BM) and ternary rhythm (TR). For POS, we rely on gold annotation from DTA and the TIGER Corpus, according to the STTS tagset. We train and test across several genres to determine the most robust POS model for our purposes. See Table 2 for an overview of the POS models. We find that training on TIGER is not robust to tag across domains, falling to around .80 F1-score when testing against different genres from DTA. Training on the whole DTA or on Belletristik (fiction/literature) is however sufficient to tag poetry or fiction. The results suggest that this model degradation across domains is mainly attributable to the curious historical orthography in DTA, and to a lesser extent due to local syntactic inversions that likely occur because of the strict metrical form (Gopidi and Alam, 2019).

Syllables	16,437,521
Tokens	11,820,483
Lines	1,607,428
Stanzas	252,841
Poems	59,424
Authors	265

Table 1: Corpus Size, Deutsches Lyrik Korpus v4

Test	Train				
	TIGER	DTA	DTA+TIGER	Belletristik	Poetry
Poetry	.795	.949	.948	.947	**.953**
Belletristik	.837	**.956**	.954	.955	.955
DTA Zeitung	.793	**.934**	.933	.911	.900
TIGER	**.971**	.928	.958	.929	.913

Table 2: Evaluation of POS taggers across genres. F1-scores.

See example (1) for an annotated line of poetry with BM and TR. BM includes binary syllable prominence (+/-) and foot boundaries ( ). TR segments the verse into rhythmic groups at caesuras (:) and in these segments allows for main accents (2), side accents (1), and no accent (0). The example line illustrates a regular iambic pentameter, a caesura at the comma, and two rhythmic groups of the same length but with different form.

(1) met="-+|-+|-+|-+|-+|" rhythm="01020:20102:"

Ge-DUCK-te H¨UT-ten, PFA-de WIRR ver-STREUT,

TR differs from BM such that TR operates top-down from rhythmic segments to find natural speech rhythm, while BM adheres more to a conventional metrical poetry analysis and starts bottom-up from syllable prominence. Annotators largely followed an intuitive notion of rhythm, and incorporated philological knowledge to consider the (schema) consistency of the poem. They were also instructed to prefer longer feet over short ones (where applicable).

Our inter-annotator agreement is substantial. Five poems were annotated by two annotators, and calculated on each syllable, Cohen κ for metric syllables was at .95, and .84 for rhythmic syllables. For the latter, mainly side accents (1) were confused. Caesuras alone had a kappa of .92. Metric feet were more challenging and will not be discussed here, as there are multiple reasons why feet boundaries can be ambiguous.

The CRF models are trained on syllabified lines without punctuation. The features contain the syllable token, including two syllables to the left and the right, and also orthographic features like capitalization. During training, we also allow it to see the surrounding labels (which are not available during testing). We achieve 95% F1-score for POS and BM on syllable level, while TR (including caesuras) still performs well with 83% F1-score. The confusion of the TR model is similar to humans who also confuse main and side accents.

2 Experiments

2.1 Stress Hierarchy

To determine the likelihood of a word belonging to a certain POS class being stressed or unstressed, we iterate over our set of 1.6 million lines of poetry, using the CRF models to annotate the Corpus for BM and POS. For our experiments, we simplify the tagset. We then count how often a POS tag falls into a metrically stressed or unstressed syllable. Multisyllabic words have mostly lexically fixed stress patterns, e.g., German words with two syllables are usually trochaic, where the first syllable is stressed and the second syllable is unstressed (the notable exceptions being alTAR and naTUR). For words with three syllables, we found that nouns are more likely to follow the (+,-,+) pattern, while verbs prefer (-,+,-). In the following, we only measure the prominence of monosyllabic words, as these display the most ambiguity, as the stress of monosyllabic words is mostly determined by their context.

Anttila et al. (2018) also determined a stress hierarchy, only for sentential stress in political speeches and not for syllables in poetic lines. They are able to establish a stress hierarchy of POS-tags, such that NOUN >ADJ >VERB >FUNC. This shows that functions words (FUNC, e.g. KONJ, ART, APPR, etc) are seldom the main accent of a sentence, while nouns are usually stressed. Based on our corpus, we determined the following hierarchy for mono- syllabic word forms:

NOUN > VERB modal > VERB full > ADJ > ADV > FUNC.

This hierarchy reflects the ratio r of stressed to unstressed syllables, normalized to 1. The ranking of all POS with their respective r can be seen in Table 3. When a POS class is equally likely to be stressed or unstressed, r will be 1.0, or 1:1. For a ratio r = 16.0 (16:1), the word class is 16 times more likely to be stressed. We found it striking that modal verbs are stressed so strongly (3.8:1). We also determined that monosyllabic verbs are more likely in metrically strong positions than monosyllabic adjectives, which differs from what Anttila et al. (2018) found. However, the end s of the hierarchy (nouns and function words) are the same. Plus, this study does not distinguish between adverbs and adjectives and also have only one verb class.

Related work of Nenkova et al. (2007) determined the strength of the POS signal for word accent prediction and developed the 'accent ratio' measure that determines the likelihood of a wordform (not POS) to be prominent or not, and showed hierarchies for word stress, while (Greene et al., 2010) looked at monosyllabic word. To the best of our knowledge, a dedicated hierarchy concerning a part-of-speech stress ratio has not been constructed yet.

2.2 Context Stress Ambiguity

As words are heavily dependent on their context regarding their stress, we look at the immediate left and right context of POS tags, i.e., which POS tag occurs next to it. For brevity, we only show the left context. We retrieve the stress ratio for particular monosyllabic word classes dependent on their context. Context words can be multisyllabic. See Table 4 for an overview of nouns, modal verbs, adjectives, adverbs and demonstrative pronouns.

We can see that the hierarchy from Table 3 reiterates for contextual dependence. If a word is preceded by a conjunction (KO), then the likelihood of stress is higher. However, nouns never lose their prominence (r > 1), regardless of context. Most interestingly, adverbs, which are quite balanced, also show a balanced context dependence, while modal verbs are still mostly stressed,

except when they are preceded by another modal verb. We acknowledge that this table can be problematic, such that some of these contexts seem atypical for particular word classes. Future research should investigate the frequency of particular contexts, and how significant they are. Lastly, the models may also introduce a systematic error that disproportionately affects certain POS classes.

2.3 Prose vs. Poetry Classification

To determine features that distinguish prose from poetic writing on a syntactic and rhythmic level, we perform classification with a regularized linear discriminant analysis that allows us to interpret feature loadings. Regularization is necessary, as many features are collinear, making the feature loadings not interpretable (as collinear features will be important for both classes, without contributing to the classification).

For features, we use POS n-grams and rhythmic groups. POS n-grams are straightforward, where only subsequent POS tag sequences are considered. For rhythmic groups, we use the CRF trained ternary rhythm (TR) with caesuras, and split the sequences at caesuras. Consequently, a line '0201:020' yields the features '0201' and '020'.

First, we tried to classify 100k sentences from poetry and prose (literature) respectively. We extracted the sentences from DTA. For POS n-grams, using unigrams or trigrams did not beat the random 50% baseline. Training the classifier on bigrams achieves 67% F1. Classifying sentences based on rhythmic groups performed at 56%, marginally better than the random baseline.

However, when we trained with POS bigrams on whole documents, F1 is around 93%, possibly because this classification relies on rare bigrams, such that sentence structure is often the same in poetry and prose, but when it differs, it differs strongly. See Table 5 for an overview of feature weights from this latter classification. Higher standing features are more important. It is notable that the positive features for prose include a significant amount of foreign language material (FM) and bigrams with a pronoun followed by 'not a word' are most discriminative for this genre.

Word Class	Label	Stress Ratio	Examples
Noun	**'NN'**	16.8	Baum, Welt, Herz, Scherz
Modal Verb	**'VM'**	3.81	muss, kann, soll
Full Verb	**'VV'**	3.67	springt, läuft, sitzt, rennt
Truncated Verb Suppl.	'TR'	3.00	an-, ab-, auf-
Auxilliary	**'VA'**	2.23	hat, ist
Adjective	**'ADJ'**	1.85	rot, hoch, rund, schnell
Named Entity	'NE'	1.44	Hans, Lutz, Max, Liz
Interjection	**'IT'**	1.29	O, Oh, Ach
Particle	**'PT'**	1.28	zu, nicht
Adverb	**'ADV'**	1.14	schon, bald, doch
Cardinal Number	'CA'	1.04	3, 5, 8, 9
Not a word	'XY'	0.85	3:7, H2O, D2XW3
Foreign Material	'FM'	0.79	A, big, fish
Reflexive/Relative Pronoun	'PR'	0.75	sich, dich, mir, der
Pronominal Adverb	'PA'	0.72	drin, dabei
Interrogative Pronoun	'PW'	0.70	wer, was, wo, wann
Indefinite Pronoun	'PI'	0.62	man, kein
Personal / Possessive Pronoun	**'PP'**	0.34	ich, er, ihm, mich, dir, eins
Preposition	'AP'	0.26	im, zur, am
Konjunction	**'KO'**	0.24	um, weil, ob, und, als, wie
Demonstrative Pronoun	'PD'	0.09	dies
Article	**'AR'**	0.06	der, die, das, ein

Table 3: POS tag stress hierarchy. Stress ambiguous monosyllaba. Ratio stressed / unstressed syllables, e.g. 16:1, 4:1, 2:1, .5:1, etc.

Nouns		Modal Verbs		Adjectives		Adverbs		Demonstr.	Pron.
Ratio	Context	Ratio	Context	Ratio	Context	Ratio	Context	Ratio	Context
42.0	KO_NN	30.33	KO_VM	8.01	KO_ADJ	3.78	AR_ADV	2.0	PD_PD
33.7	ADJ_NN	15.85	PD_VM	4.17	PT_ADJ	3.05	KO_ADV	0.75	PA_PD
31.0	AR_NN	10.31	VV_VM	3.44	PW_ADJ	2.19	PD_ADV	0.67	AR_PD
24.6	PD_NN	5.71	AR_VM	3.15	PR_ADJ	1.78	PI_ADV	0.62	KO_PD
21.0	AP_NN	5.54	PP_VM	2.98	PP_ADJ	1.75	PP_ADV	0.5	FM_PD
17.7	PP_NN	4.44	VA_VM	2.82	ADV_ADJ	1.64	PR_ADV	0.42	PI_PD
16.5	PR_NN	4.42	PI_VM	2.35	PI_ADJ	1.54	AP_ADV	0.36	PR_PD
16.0	FM_NN	4.36	PW_VM	2.23	VV_ADJ	1.45	XY_ADV	0.29	PW_PD
14.0	PI_NN	4.28	ADJ_VM	2.15	VA_ADJ	1.43	FM_ADV	0.29	CA_PD
9.91	NN_NN	3.5	FM_VM	1.79	NN_ADJ	1.24	PW_ADV	0.26	PP_PD
9.56	PW_NN	3.33	XY_VM	1.64	FM_ADJ	1.04	ADJ_ADV	0.24	ADJ_PD
8.94	VA_NN	2.96	PA_VM	1.62	PA_ADJ	1.02	ADV_ADV	0.18	IT_PD
8.94	PT_NN	2.58	NE_VM	1.49	PD_ADJ	0.99	NN_ADV	0.17	NN_PD
8.01	ADV_NN	2.21	ADV_VM	1.15	AR_ADJ	0.84	VV_ADV	0.17	ADV_PD
7.03	VV_NN	1.41	NN_VM	1.0	IT_ADJ	0.77	NE_ADV	0.15	VV_PD
6.57	IT_NN	1.28	PT_VM	0.88	VM_ADJ	0.76	PA_ADV	0.15	AP_PD
6.56	NE_NN	1.25	AP_VM	0.75	AP_ADJ	0.73	VM_ADV	0.14	VM_PD
4.12	VM_NN	1.12	PR_VM	0.61	ADJ_ADJ	0.67	VA_ADV	0.13	XY_PD
3.83	XY_NN	1.0	IT_VM	0.36	XY_ADJ	0.6	CA_ADV	0.11	VA_PD
2.83	PA_NN	0.17	VM_VM	0.24	NE_ADJ	0.54	PT_ADV	0.11	NE_PD
1.67	CA_NN			0.12	CA_ADJ	0.41	IT_ADV	0.05	PT_PD

Table 4: Context dependence of monosyllabic word stress.

Prose Feature Weights		Poetry Feature Weights	
-8611	PWS_XY	21821	PRF_KOKOM
-7983	PWAV_KOKOM	20646	PAV_KON
-7907	PPER_FM.en	19429	KOUI_PRF
-7251	FM.la_APPR	17877	APZR_VAFIN
-6932	FM_PAV	15542	APPO_VVFIN
-6634	VVIZU_PTKNEG	15330	VAPP_ADV
-6220	XY_PIS	14724	VMFIN_PTKVZ
-5919	VAIMP_APPR	14487	NN_PTKANT
-5483	ADJD_FM.la	14256	KOUI_VVIZU
-5267	FM_PDAT	13408	PTKNEG_CARD

Table 5: Feature Loadings of Prose vs. Poetry, document classification with .93 F1

	Sentences	Documents
Rhythm	.56	.84
POS 2-grams	.67	.93
POS 1-/3-grams	.50	-

Table 6: Poetry vs. Prose Classification F1-micro

2.4 Versemeter vs. Enjambement

Our setup also allows us to get an impression of the interaction of enjambement with verse measures and also POS transitions between lines. Enjambement is an integral part of many poetic lines. It typically signifies incomplete syntax at the end of a line, such that the end of the line encourages a pause in speech, but the sentence, or clause, or phrase, or word is not yet finished. We use a simple way to operationalize

enjambement, by assigning enj+ to lines that do not end on a punctuation mark, and enj- to lines that do. Beyond obvious cases (ART NN does not cross punctuation/clause boundaries), we could not identify clear preferences of enjambement for particular POS transistions.

We implement a set of regular expressions to detect a broad set of verse measures based on the syllable prediction of the meter CRF (here, we use I for stressed and o for unstressed syllables, while the symbol before '?' is optional and '$' is the end of the line). Unsurprisingly, we find that lines with fewer stressed syllables prefer enjambement more. However, for measures with six stressed syllables, the 'running measure' hexameter (Ioo?Ioo?Ioo?Ioo?IooIo$) is more sympathetic with enjambement than any other measure with a probability of p(enj+) = .41, while the alexandrine (oIoIoIoIoIoIo?$) dislikes it, ranking as most unlikely with p(enj+) = .16, compared to all other measures.

3 Conclusion

We have shown experiments on the intersection of syntax and speech rhythm, outlining stress hierarchies with and without context, questioning previous research. Lastly, we have shown that a classification of documents on POS bigrams shows clear distinctive features of poetry vs. prose, while classifying sentences is challenging. In the end, we also discussed first explorations regarding enjambement and its interaction with syntax and different verse forms.

References

Arto Anttila, Timothy Dozat, Daniel Galbraith, and Naomi Shapiro. 2018. Sentence stress in presidential speeches. Lingbuzz Preprints.

Derek Attridge. 2014. The rhythms of English poetry. Routledge.

Valentin Wagner Matthias Schlesewsky Blohm, Stefan and Winfried Menninghaus. 2018. Sentence judgments and the grammar of poetry: Linking linguistic structure and poetic effect. Poetics, 69:41–56.

Amitha Gopidi and Aniket Alam. 2019. Computational analysis of the historical changes in poetry and prose. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 14–22.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In Proceedings of the 2010 conference on empirical methods in natural language processing, pages 524–533.

Thomas Haider and Steffen Eger. 2019. Semantic change and emerging tropes in a large corpus of new high german poetry. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 216–222.

Willem JM Levelt. 1993. Speaking: From intention to articulation, volume 1. MIT press.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Thomas Jacobsen, and Christine A Knoop. 2017. The emotional and aesthetic powers of parallelistic diction. Poetics, 63:47–59.

Winfried Menninghaus, Valentin Wagner, Christine A Knoop, and Mathias Scharinger. 2018. Poetic speech melody: A crucial link between music and language.PLoS One, 13(11):e0205980.

Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky. 2007. To memorize or to predict: Prominence labeling in conversational speech.

Dynamic Systems for Humanities Audio Collections: The Theory and Rationale of Swallow

Jason Camlot
jason.camlot@concordia.ca
Concordia University, Canada

Tomasz Neugebauer
Tomasz.Neugebauer@concordia.ca
Concordia University, Canada

Francisco Berrizbeitia
rancisco.berrizbeitia@concordia.ca
Concordia University, Canada

Keywords: Metadata, Systems, Data
Topics: English, North America, Contemporary, digital archiving, metadata standards, systems, and methods, Library & information science, Literary studies

Dynamic Systems for Humanities Audio Collections: The Theory and Rationale of Swallow

Jason Camlot, Tomasz Neugebauer, Francisco Berrizbeitia

This paper approaches a system that has been designed, and continues to be in development, for the aggregation of metadata surrounding collections of documentary literary sound recordings, as an

object for theoretical and practical discussion of how information about diverse collections of time-based media should be managed, and what such schema and system development means for our engagement with the contents of such collections as artifacts of humanist inquiry. Swallow (Swallow Metadata Management System 2019), the interoperable spoken-audio metadata ingest system project that is the boundary object for this talk, emerged out of the goals of the SpokenWeb SSHRC Partnership Grant research network to digitize, process, describe, and aggregate the metadata of a diverse range of sound collections documenting literary and cultural activity in Canada since the 1950s. Our talk, collaboratively written and delivered by a literary scholar and critical theorist, a digital projects and systems development librarian, and a library developer / programmer, outlines 1) a theoretical rationale for the audiotext as a significant form of data in the humanities, 2) consequent modes of description deemed necessary to render such data useful for humanities scholars, and 3) a rationale for the development of a specific form of database system given the material and systems contexts that inform our national holdings of documentary literary sound recordings at the present time.

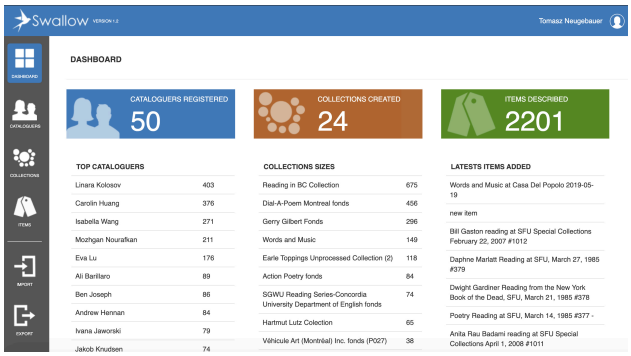


Figure 1. Screenshot of Swallow Dashboard

Rationale of digitized audiotexts as humanities research data

Why is the study of literary recordings important? With the introduction of viable sound recording technology in the last decade of the nineteenth century, a new form of material data for literary analysis emerged: the audiotext. This new literary artifact, based on the sound signal of the poet's acoustic performance, has been defined as "a semantically denser field of linguistic activity" that demands new methods of analysis (Bernstein). Once digitized, engagement with sound recordings move us further down the path of understanding "The Text" as "a methodological field" (Barthes), and towards increasingly elaborate "rationales of audio text" (Clement) realized in relation to information systems for the purposes of software analysis, content modeling, and cataloguing. The digitized archive of voices invites the production of thorough metadata, and thus expands the ways in which we may remodel our understanding of historical cultural artifacts (Camlot 2019, Camlot and Mitchell, Gitelman). Over the past two decades, due in part to the founding of online repositories of literary recordings like PennSound (PennSound center for programs in contemporary writing at the University of Pennsylvania 2020) and SpokenWeb (SpokenWeb 2020) the potential of the literary sound archive for research has become discernible. The implications of such research are momentous for future methodologies in the humanities, and yet the realization of such research is limited, at present, by the status and structure of our collections of literary sound recordings. In Canada, thousands of hours of literary readings, performance and related activities have been recorded on magnetic tape since the 1950s. The affordances of tape recording and subsequent media technologies enabled new literary uses of sound recording, including the documentation of literary events, readings and conversations, both public and private. This extended the reach of capturing literary forms and events, and consequently, transformed our understanding of what comprised the literary and cultural events. As valuable archival materials, these recordings represent a massive, largely untapped, undifferentiated, and often undiscovered resource for the study of literary art, culture and society. Research in literary sound studies depends upon a collaborative, interdisciplinary approach to the development of historical knowledge about these valuable cultural heritage materials, and of a networked structure of digital repositories and research-focused interfaces for the preservation and meaningful presentation of these materials to researchers, students, artists and the public. Swallow, the metadata system we launched in 2019 and intend to augment into a robust tool for sharing metadata about our national audio holdings across multiple institutions and a diverse range of metadata and access systems, is a coded software system that represents and works to realize the theoretical idea of audiotexts as significant digital humanities data.

A Poetics of Audio Description, or, a Metadata Schema for Literary Audio

The diversity of recordings held at SpokenWeb partner institutions across Canada include recorded conversations, dictations, compilations, performances, interviews and lectures, among numerous other generic audiotextual categories. How do we begin to describe

the features and contents of such recordings, and in what ways may we make collections of documentary sound recordings speak to each other, beyond their local venues of origin and siloed media formats and archives?

Metadata and its structure has the potential to open and limit avenues of research. The aggregation of metadata from partner institutions requires either choosing an existing access system, such as Islandora (Islandora Open source digital asset management 2020) or Avalon (Avalon Media System 2020), with its own metadata schema, or to design one based on the evolving research needs of the project. We have chosen the latter because we do not wish to limit our research questions to any single digital repository platform. During the extensive metadata collection phase of our research program, we are seeking maximum usability. We are pursuing our work in describing the assets and contents of collections of documentary literary sound recordings using an agile, iterative development process. We have developed our long-form metadata schema over the course of a year through the research, discussion, and regular meetings of a SpokenWeb Metadata Task Force consisting of librarians, literary researchers and students from across our partnership. We have provided the metadata schema (Camlot et al. 2020) and ingest software out to student cataloguers quickly for testing and feedback, so that the schema has been shaped by research use from the outset, and will continue to be shaped by researcher feedback throughout this seven-year cataloguing project. We based the SpokenWeb schema on standards such as IASA (International Association of Sound and Audiovisual Archives 1999), AACR2 (Joint Steering Committee for Revision of AACR 2005) and MODS (Library of Congress 2018), focusing on the most useful simplified instructions for the audiotext metadata context.

The initial result was a SpokenWeb Metadata Schema with standard core fields such as: Title, Rights, Creators/Contributors, Production Context, Genre, and Contents. The material artifact description recommendations in the schema include detailed sub-fields reflecting research interest, such as: AV Type, Material Designation, Physical Composition, Storage Capacity, Extent, Playing Speed, Track Configuration, Playback Mode, Tape Brand. The digital file description was a multiple field from the beginning, allowing for multiple files per material artifact. We later added Contents and Notes sub-fields that are capable of storing XML markup to Digital File Description, to make it possible to associate content annotations for each of the files in case of multi-file items. Following feedback from cataloguers who encountered event recording spanning multiple artifacts, it made sense to make Material Description into a multiple field as well. Other examples of changes include: the addition of "Performance Date" to Date Types and "Classroom recording" to Production Context.We recognized the research need to catalogue event/venue names and locations early on, and included location subfields for venue names, latitude and longitude, links to OpenStreetMap nodes, and notes for additional context.

For qualifying and describing authorship, we predefined a controlled vocabulary by selecting from the Library of Congress relators those roles relevant to audiotext, such as interviewer, author, presenter. In the second iteration of the schema, we made it possible to associate multiple roles for a contributor, and following feedback from one of our partner institutions working with First Nations content, we added the role of "Elder". In addition, soon after the launch of Swallow, the First Nations Metis and Inuit Indigenous Ontology was released, and we added the Nation subfield and controlled vocabulary lookup based on this for Creators/Contributors.

The Contents field has been designed to facilitate navigation, searching and linking audio content of one recording to a wide collection of others. The inclusion of a Related Works field is one way to facilitate linking. The content information needs to be able to include time stamped structural and descriptive information about the works and creators on the recordings. There are many possible tools and formats to perform this type of time stamped metadata. Swallow allows for the storage of contents field information using text or XML markup. We also developed recommendations for how to annotate literary audio for the project and a Python/Flask script to convert text formatted annotations to Avalon XML (Neugebauer 2020). Avalon's XML format for audio annotations was extended slightly to make it possible to package multiple digital files' annotations into a single Avalon XML file.

Swallow and the rationale for an iterative database design

The SpokenWeb collections are diverse and dispersed in archives and different repository systems. The networked approach we have pursued has allowed the research network to proceed with aggregating the metadata from partners without having to limit the infrastructure by choosing only one repository system with its associated metadata schema. However, we have also needed to move beyond spreadsheets, to improve the accuracy and usability of the cataloguing work. A spreadsheet of metadata quickly becomes difficult to manage as the number of columns increases to accommodate multiple fields, to catalogue multiple creators/contributors, for example. In addition, the

research requirements of the project call for a highly customizable and extendible set of metadata fields that will change over time as we analyze and catalogue the content and learn more about the collections. We needed to be able to rapidly iterate new versions of interfaces for evolving metadata schema. Our solution has been to avoid hardcoding the metadata schema into the relational database, and instead to store the metadata as unstructured JSON inside the database. A layer of abstraction then allows us to quickly modify configuration files that define the metadata fields and their associated controlled vocabularies and lookups. The resulting system's cataloguing interface may then be changed quickly, and is able to accommodate multiple schemas. Indeed, we have implemented numerous additions and changes to the metadata system over the course of our first year of cataloguing in response to the nature of specific collections we are describing.

Swallow makes it possible for literary audio holding institutions to aggregate their metadata in a single system. The required metadata fields for access and preservation can change over time and vary by document type and source institution. Swallow allows for the aggregation of metadata in multiple metadata schemas in the same system. It generates a schema-specific cataloguing interface for each item stored, based on the item's respective schema configuration files that include a list of steps, metadata fields and controlled vocabularies. The Swallow schema specification functionality currently allows for the inclusion of URIs alongside metadata values that are a part of the SpokenWeb Schema, such as links to VIAF, OpenStreetMap, Wikidata, and Rights Statements.

Future development

Figure 2 shows the proposed flow of information with Swallow at the center acting as a central metadata repository. The functionality that we are proposing to develop includes the capacity to batch ingest metadata from other systems currently used by SpokenWeb partners into the latest version of the SpokenWeb schema in Swallow. The functionality to access batch import mappings on the interface exists, and we will develop the mappings for Islandora, Avalon, and AtoM in collaboration with partner institutions, Concordia Spectrum Research Repository, SFU Radar, University of Alberta Dataverse, University of Toronto (Dataverse), UBC Brain Circuits (FRDR). The proposed functionality includes the development of batch editing that would allow cataloguers to modify the metadata of collections of items. We will also develop mapping specifications that will allow for the migration of items catalogued in one schema version into another. The proposed development would build into the interface the functionality to lookup and export metadata from Swallow to Wikidata. In addition, we will develop mappings for export of collections of metadata as data from Swallow to DataVerse and FRDR (Canadian Association of Research Libraries & Compute Canada 2020) .

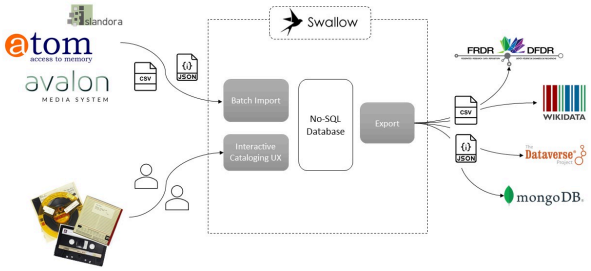


Figure 2. Swallow as central metadata repository

Conclusion

Our work in metadata and systems development can be read as the development as a poetics of the audiotext, that is a poesis (a making) of the descriptive and digital infrastructure that allows digitized audio artifacts – artifacts that capture the sound of literature in its social contexts during the postwar era – to survive and take flight in the imaginations of contemporary critics, scholars, poets and listeners.

Bibliography

Avalon Media System 2020, viewed June 15, 2020, https://www.avalonmediasystem.org/.

Bernstein, Charles, ed. Close Listening: Poetry and the Performed Word. New York, NY: Oxford UP, 1998.

Barthes, Roland. "From Work to Text." Trans. Richard Howard. The Rustle of Language. New York: Farrar, Straus and Giroux, 1986. 56-64.

Camlot, Jason. Phonopoetics: The Making of Early Literary Recordings. Stanford, CA: Stanford University Press, 2019.

Camlot, Jason and Christine Mitchell. "The Poetry Series." Amodern 4 (March 2015). http://amodern.net/article/editorial-amodern-4/. Web.

Camlot, J., Dowson, R. Ferrier, Fong D., Kail R., Luyk S., Meza A., Neugebauer T., Wiercinski J., Barillaro A., Hannigan L., Lu E., Mash C., Pickering H , Chandler N., Knudsen J, Kolosov L, Roberge, A. SpokenWeb Metadata Scheme and Cataloguing Process 2020, viewed June 15, 2020, .

Clement, Tanya. "Towards a Rationale of Audio-Text." Digital Humanities Quarterly 10.2 (2016). Web.

Federated Research Data Repository (FRDR) 2020, Canadian Association of Research Libraries & Compute

Canada, viewed June 15, 2020, https://www.frdr-dfdr.ca/repo/.

Gitelman, Lisa. Always Already New: Media, History and The Data of Culture.Cambridge, MA: MIT Press, 2008.

"The IASA Cataloguing Rules (IASA 1999)". The International Association of Sound and Audiovisual Archives, viewed June 15, 2020, https://www.iasa-web.org/cataloguing-rules/.

Islandora Open source digital asset management 2020, Islandora Foundation, viewed June 15, 2020, https://islandora.ca/.

Joint Steering Committee for Revision of AACR (2005). Anglo-American cataloguing rules, 2002 revision. Chicago, Ill, American Library Association.

Library of Congress 2018. Outline of Elements and Attributes in MODS Version 3.7. viewed June 15, 2020. https://loc.gov/standards/mods/mods-outline-3-7.html#name,%20v.3.

Neugebauer, Tomasz. Python Script to Convert Audio Annotation Files. 2020, GitHub, viewed June 15, 2020, https://github.com/photomedia/AudioAnnotateConvert .

PennSound center for programs in contemporary writing at the University of Pennsylvania 2020, viewed June 15, 2020, http://writing.upenn.edu/pennsound/.

RIGHTS STATEMENTS 2018, viewed June 15, 2020, https://rightsstatements.org/page/1.0/?language=en.

SpokenWeb 2020, viewed June 15, 2020, https://spokenweb.ca/.

Swallow Metadata Management System 2019, GitHub, viewed June 15, 2020, https://github.com/spokenweb/swallow.

VIAF: The Virtual International Authority File 2020. viewed June 15, 2020, http://viaf.org/.

The Digital Curation of Broadcasting Archives at the Canadian Broadcasting Corporation: Curation Culture and Evaluative Practice

Asen Ivanov
ivanova@uoguelph.ca
Michael Ridley Postdoctoral Fellow in Digital Humanities, THINC Lab, University of Guelph

Keywords: broadcasting archives, digital curation, evaluation, information practice, practice theory
Topics: English, North America, Contemporary, cultural analytics, digital archiving, Library & information science, Media studies

The Digital Curation of Broadcasting Archives at the Canadian Broadcasting Corporation: Curation Culture and Evaluative Practice

Asen O. Ivanov
Michael Ridley Postdoctoral Fellow
University of Guelph
DH 2020

This paper presents a doctoral project on the digital curation practices at the news archive of the Canadian Broadcasting Corporation (CBC) (Ivanov 2019). The CBC archive is the largest news media archive in Canada, with over 500,000 hours of digital archival holdings dating back to the 1950s. Data for the project were collected through an organizational ethnography at the CBC news archive's headquarters in Toronto, Canada ([FN1]). The project examined how archivists at the CBC evaluate, transform, enrich, and reuse archival news materials and sought to understand how organizational and computational processes and systems intersect with human factors in digital curation practice ([FN 2]). This paper describes the project's approach and summarizes its main findings and conclusions.

Broadcasting news archives provide rich, sometimes visceral, representations of the past. This makes them a vital element of broadcasting journalism's visual language, where they serve as "mnemonic devices" for reconstructing the past (Conboy 2012). But news archives are more than storerooms of information. They are products of a time and place, created through the articulation of different kinds of practices and technologies. As archival ethnographies demonstrate, rather than mere storerooms, archives are more accurately described as dynamic sites of knowledge-making in which different forms of knowledge about the past are configured through information practices such as selection, classification, and organization (Gracy 2004).

The study presented in this paper amplifies the semiotic and material dimension of the digital curation practices at the CBC news archive. It does so by taking up an approach developed for the study of knowledge-making in the sciences and applying it to the curation of the CBC archive, positioning the latter as a sub-set of "social knowledge making" (Camic, Gross, and Lamont 2011). The fundamental assumption of this approach is that neither the social world of actors and

actions, the material world of objects and technologies, nor the cultural world of meanings and symbols could be coherently organized or function independently from one another. They are all intertwined in what Barad (2003) calls practices of matter and meaning through which the world comes "to matter." One way to understand these three worlds is to examine the overlapping social, material, and cultural arrangements and mechanisms that bind them together within the context of historically situated expert systems of knowing (Knorr Cetina 1999). In less abstract terms, then, using this approach asks the analyst to 'open' practices of knowledge-making, read and organize their contents and symbols, and analyze how they are used.

Following prior work in this analytical tradition, in my dissertation, I analyzed the organizational structures and social interactions at the CBC news archive and the information systems and tools of curation (Latour 1987). I also examined how cultural and symbolic elements such as rules, norms, beliefs, and codes provide strategies for action that allow archivists at the CBC to address pragmatic problems in their immediate environment (Swidler 1986). Lastly, I examined how information is evaluated through the digital curation lifecycle at the CBC news archive, focusing on the methods of comparison, criteria, conventions, and tools supporting these evaluations (Beljean, Chong, and Lamont 2016).

What I found is that the organizational structure of the CBC news archive resembles the axioms of a networked organization, similar to other contemporary knowledge organizations (e.g., software firms). This work environment is characterized by interactions within and across teams and constant information and knowledge exchange. The CBC archivists are expected to develop expertise on the job through an apprenticeship and are hired only in terms of their suitability for this type of work. When hiring new employees, the management seeks for either one of two signals of suitability: training in broadcasting journalism (demonstrated by work experience or a graduate degree) or information management (demonstrated by a library and archival graduate degree). The rationale, I was told, is that the former group understands what makes good journalism, and the latter how to select, classify, catalogue, and organize information. Once on the job, the archivists are expected to master both- they need to learn how to recognize the codes and conventions of journalism and how to manage data and information in complex information systems. It is already in these rudimentary job requirements that the semiotic and material dimensions of digital curation come to analytical light.

Curation practices at the CBC have been evolving gradually for decades since the founding of the archive in the late 1930s, but established ways of working were upended by the organization's digital transformation in 2008. One consequence of digitization is that news archivists took on a more proactive role in the production of news. They become embedded into the newsroom to anticipate, rather than react to, editorial needs. Digitization accelerated the transfer of information, and consequently, the pace at which work is experienced, creating new pressures and demands for accuracy. Videotapes (as media carriers) were replaced by a complex set of digital assets, some of which are highly unstable and could be easily altered. This made it more difficult to determine and document the identity and integrity of news materials, as doing so required a new set of evaluative criteria and practices ([FN 3]).

The digital transformation of the CBC also led to the introduction of new information and media management systems, whose adoption resulted in unanticipated changes to due process at the workplace (Gerson and Star 1986). This is because while the new information systems archivists had to learn how to use are state-of-the-art, they are designed with the goals of media production in mind. The "affordances" of these systems presented material and semiotic challenges to the established ways of conceptualizing, representing, and doing curation at the CBC (Gibson 2014 [1979]). As my analysis suggests, the material challenges are addressed through information systems workarounds (Alter 2014). The semiotic challenges, on the other hand, are addressed by developing home-grown file classifications systems that map onto organizational structures and function as sense-making mechanisms that represent the spatial temporal, and causal relatedness among different organizational actors and the archival holdings (Weick 1995).

Lastly, I interviewed my research participants to learn more about the evaluative practices in their work, asking them questions about the various forms of decision-making strategies and heuristics they apply in the selection, cataloguing, and retrieval of archival news materials. This analysis, most broadly, indicates that evaluative practices are consequential during all stages of the curation lifecycle but are particularly salient at the stages of acquisition (when value is established), cataloguing (when value is enriched), and reuse (when value is negotiated to fit into a new context).

Beyond that, my interview data suggest that at the CBC news archive evaluation occurs within both situated and institutional frameworks of understanding in practice (Ivanov 2020). Archivists base their evaluations of archival news materials in the CBC

archival policy-which succinctly outlines three value criteria: cultural-value, significance to the corporation, and reuse-as well as in the codes and conventions of broadcasting journalism (which often entails making aesthetics judgements about the content and visual quality of news materials). But as my analysis further suggests, when the archival policy and the codes and conventions of broadcasting journalism cannot provide concrete strategies for action, archivists draw on their subjective experiences, taste, and preferences in evaluation. They, thus, base their evolutions on what institutions identify as valuable but equally so on what they, based on their subjective experiences, recognize as valuable. The intertwinement of institutional and situated frameworks in evaluation differs depending on the situation, the archivist doing the evaluation, the type of materials evaluated, and the goals of evaluation (e.g., for acquisition, cataloguing, or reuse), but it consistently recurs. The outcomes of these evaluations are inscribed in the materiality of the archive in various ways as they are written into item- and collection-level metadata records, taxonomies, indices, and classification systems.

In conclusion, based on these findings, I propose that aside from being supported by socio-technical systems, digital curation practices are also supported by cultural systems. These cultural systems sustain collective action and situated cognition by providing vocabularies of motives and strategies for action that cohere in a curation culture. Second, I propose that evaluative practices are an integral part of digital curation and a prerequisite for enhancing the value of digital materials. They are imbricated in institutional and situated frameworks of understanding, and while computationally imperfect are practically indispensable. It is in the context of evaluation, I thus argue, that human intelligence and judgment most saliently influence the semiotic and material outcomes of digital curation practice.

FOOTNOTES:

[FN1] - For my methodology see Ivanov (2019, 98-107); See also summary in Ivanov (2020, 2892-2893).

[FN2] - Digital curation is defined as "the active management of data or digital materials through their lifecycle with the goal of ensuring their preservation, enhancing their value, and facilitating their use and reuse" (See, Dallas 2016, 430; See also, Higgins 2008).

[FN3] - In archival science, the evaluation of the integrity and identity of archival records are key perquisites for determining their authenticity and reliability (Duranti 1995). These concerns translate to digital curation, where have been further elaborated as concepts and practices (For discussion see Duranti and Thibodeau 2006; Ross 2012; Becker 2018; Ivanov 2019, 32-43).

REFERENCES

Alter, Steven. 2014. "Theory of Workarounds." Communications of the Association for Information Systems 34:1041-1066.

Barad, Karen. 2003. "Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter." Signs: Journal of Women in Culture and Society 28 (3): 801-831. https://doi.org/10.1086/345321

Becker, Christoph. 2018. "Metaphors we Work by: Reframing Digital Objects, Significant Properties, and the Design of Digital Preservation Systems." Archivaria 85: 6-37.

Beljean, Stefan, Philippa Chong and Michèle Lamont. 2016. "A Post-Bourdieusian Sociology of Valuation and Evaluation for the Field of Cultural Production," In Routledge International Handbook of the Sociology of Art and Culture, edited by Laurie Hanquinet and Mike Savage, 38-48. New York: Routledge.

Camic, Charles, Neil Gross, and Michèle Lamont. 2011. "The Study of Social Knowledge Making." In Social Knowledge in the Making, edited by Charles Camic, Neil Gross, and Michèle Lamont, 1-43. Chicago, IL: The University of Chicago Press.

Conboy, Martin. 2012. "How Journalism Uses History." In How Journalism Uses History, edited by Martin Conboy, 1-15. New York: Routledge.

Dallas, Costis. 2016. "Digital Curation Beyond the 'Wild Frontier': A Pragmatic Approach," Archival Science, 16 (4): 421-457. https://doi.org/10.1007/s10502-015-9252-6

Duranti, Luciana. 1995. "Reliability and Authenticity: The Concepts and their Implications." Archivaria 39: 5-9.

Duranti, Luciana, and Kenneth Thibodeau. 2006. "The concept of record in interactive, experiential and dynamic environments: The view of InterPARES." Archival Science (6)1: 13-68.

Gerson, Elihu M., and Susan Leigh Star. 1986. "Analyzing Due Process in the Workplace." ACM Transactions on Information Systems 4 (3): 257-270.

Gibson, James J. (2014 [1979]) "The Theory of Affordances." In The People, Place, and Space Reader, edited by Jen Jack Gieseking, William Mangold, Cindi Katz, Setha Low, Susan Saegert, 56-61. New York, Routledge.

Gracy, Karen F. 2004. "Documenting Communities of Practice: Making the Case for Archival Ethnography." Archival Science 4 (3-4): 335-365. https://doi.org/10.1007/s10502-005-2599-3

Higgins, Sarah. 2008. "The DCC Curation Lifecycle Model," International Journal of Digital Curation 3 (1): 134-140. https://doi.org/10.2218/ijdc.v3i1.48

Ivanov, Asen O. 2019. "The Digital Curation of Broadcasting Archives at the Canadian Broadcasting Corporation: Curation Culture and Evaluative Practice." PhD diss., University of Toronto. http://hdl.handle.net/1807/95833

---.2020. "Evaluative Practices in a Broadcasting Newsroom Archive: Culture, Context, and Understanding in Practice." International Journal of Communication 14. https://ijoc.org/index.php/ijoc/article/view/11591

Knorr Cetina, Karin. 1999. Epistemic Cultures: How the Sciences Make Knowledge. Cambridge, MA: Harvard University Press.

Latour, Bruno. 1987. Science in Action: How to Follow Scientists and Engineers Through Society. Cambridge, MA: Harvard University Press.

Ross, Seamus. 2012. "Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries." New Review of Information Networking 17 (1): 43-68.

Swidler, Ann. 1986. "Culture in Action: Symbols and Strategies." American Sociological Review 51 (2): 273-286.

Weick, Karl E. 1995. Sensemaking in Organizations. Thousand Oaks, CA: SAGE.

Nanohistory – An Experimental Digital History Methodology

Matthew Milner
mmilner@mun.ca
Memorial University of Newfoundland, Canada

Keywords: History, Event Modeling, Historical Method, Digital History, Networks
Topics: Global, Europe, English, North America, 15th-17th Century, 19th Century, 20th Century, data modeling, digital research infrastructures development and analysis, History

Nanohistory is a prototype digital history platform (https://www.nanohistory.org). It envisages the writing of history-as-data, using the smallest statements of historical action - nanohistories - which are documented as named graphs or 'quads'. The prototype is built on a synthesis of historiographical and information science theories focused on the role narrative and discourse play in recounting and studying the past. This paper will detail preliminary assessment of the prototype as a viable tool for historical research by a group of diverse 11 digital history projects over 2019-20. Funded by the Social Sciences and Humanities Research Council of Canada, the purpose of the assessment is to discern whether the prototype's synthesis of historical, network, and digital theories offers a viable digital historical methodology for documenting complex historical phenomena ranging from large-scale events, movement of people and artifacts, and discourse surrounding identities including gender, indigeneity, and religious culture.

Panels

Understanding Rhyme Through Network Analysis

Natalie M. Houston
Natalie_Houston@uml.edu
University of Massachusetts Lowell

Petr Plecháč
plechac@ucl.cas.cz
Institute of Czech Literature, Czech Academy of Sciences

Pablo Ruiz Fabo
ruizfabo@unistra.fr
LiLPa, Université de Strasbourg

Helena Bermúdez Sabel
helena.bermudezsabel@unil.ch
Université de Lausanne

David Birnbaum
djbpitt@gmail.com
University of Pittsburgh

Elise Thorsen
enthorsen@gmail.com
Novetta

Keywords: rhyme, network analysis, machine learning, stylometry, intertextuality
Topics: South America, Europe, English, North America, 15th-17th Century, 19th Century, 20th Century, network analysis and graphs theory and application, text mining and analysis, Linguistics, Literary studies

Understanding Rhyme Through Network Analysis

Rhyme enacts numerous relationships in poetic texts: relationships between words that share similar sounds; relationships between lines of verse that end in rhyming words; and relationships between the sound and semantic meaning of words that are linked together through rhyme. This panel brings together different approaches to using network analysis to understand the relationships that rhyme enacts in poetry in different language traditions. As the papers in this panel suggest, as one moves from considering rhyme's function within a single poem to examining larger datasets, one can also consider how rhyme connects words, documents, and/or authors within a corpus. By juxtaposing papers focusing on poetry in Czech, English, Russian, and Spanish, this panel highlights the fact that rhyme is defined differently in different linguistic and poetic contexts, due to the levels of inflection present in different languages and to the development of different poetic traditions. Some rhyme definitions focus on the shared stressed vowel of single syllables, others encompass polysyllabic rhyme, and others focus on component phonemes, rather than whole words. Thus computational approaches to rhyme must be tailored to the particular languages of the texts under study. The papers on this panel use different network analysis and graph visualization methods to examine rhyme at the level of the corpus or dataset, rather than the individual poem, in order to understand how rhyme practice changes over time, across languages, and in relation to literary canon formation. It thus contributes both to computational poetics and distant reading methodologies within the digital humanities.

Distant Reading Nineteenth-Century British Poetry With Rhyme Networks

Natalie M. Houston
University of Massachusetts Lowell

1. Introduction

The expectations and assumptions that nineteenth-century English readers brought to their reading of poetry was necessarily different from that which readers today bring to the same texts. One feature of that historical difference was their familiarity with poetic rhyme, and the assumptions about poetic language that it created. By examining rhyme words and sounds in a large dataset of English poems, we can better understand how rhyme shaped poetic discourse in the nineteenth century. This paper suggests that network analysis methods are useful for understanding the semantic networks created by the relatively limited set of rhymes available in English; for examining chronological and aesthetic explanations for changes in rhyme practice; and for exploring the relationships between poems that use the same rhyme pairs. Such analyses reveal the semantic and sonic features of conventional nineteenth- century poetry, and can thereby also distinguish unconventional or distinctive uses of rhyme.

2. Context

The vast majority of English poems published in the nineteenth century were rhymed (95% of the 108,842 poems in the Chadwyck-Healey English Poetry corpus used for this study). Both poets and readers thus expected poetry to be rhymed (McDonald 2012, 7) and many of the rhyme sounds and rhyme words used in nineteenth- century poetry were so frequently used as to create a set of implicit conventions of poetic discourse. Uncovering such implicit conventions can help reveal the structures of the field of poetry at the large scale

(Bourdieu 1993).

3. Method

This project identifies rhyme words, groups, and syllables through a method that operationalizes the historical rules for rhyme found in nineteenth-century British rhyme dictionaries, in order to match words according to historical pronunciation and poetics (Houston 2016, 2019). Three different kinds of network analyses are then performed: a rhyme word co-occurrence network, a rhyme pair co-printing network, and a textual coupling network (Houston 2017).

4. Rhyme Networks

Because rhyme word frequency in British poetry follows a power law distribution, in which a small number of rhyme words are very frequently used, followed by a long tail of additional words, the rhyme word co-occurrence network can reveal the relationship between those frequencies and the clusters of rhyme words that are most likely to occur within the same poem. The rhyme pair co-printing network, based on co-citation analysis, links specific rhyme pairs if they appear in the same poem. Together these two networks reveal the semantic and sonic patterns that structured nineteenth-century poetic discourse. The textual coupling network, based on bibliographic coupling, links two poems if they use the same rhyme pairs. This network reveals chronological and aesthetic subgroups, suggesting how rhyme practice changed through the century.

Works Cited

Bourdieu, P. (1993). "The Field of Cultural Production, or: The Economic World Reversed." In The Field of Cultural Production: Essays on Art and Literature. Ed. Randal Johnson. New York: Columbia University Press. 29-73.

Houston, N. (2019). "An Evaluation of Rhyme Detection Using Historical Dictionaries." Digital Humanities 2019, Utrecht University, Netherlands.

Houston, N. (2016). "Exploring the Rules of Rhyme: Operationalizing Historical Poetics." Digital Humanities 2016, Krakow, Poland.

Houston, N. (2017). "Measuring Canonicity: a Network Analysis Approach to Poetry Anthologies." Digital Humanities 2017, McGill University, Montreal Canada.

Jauss, H.R. (1970). "Literary History as a Challenge to Literary Theory." Trans. Elizabeth Benzinger. New Literary History 2.1: 7-37.

McDonald, P. (2012). Sound Intentions: the Workings of Rhyme in Nineteenth-Century Poetry. Oxford: Oxford University Press.

Recurrences of rhymes in 19th century Czech poetry as compared to English, German, Spanish, and Russian

Petr Plecháč

Institute of Czech Literature, Czech Academy of Sciences

In any large enough body of poetic texts rhyme pairs inevitably reoccur (cf. Reddy-Knight 2011). The talk employs network analysis in order to explore to what extent rhyme pairs are shared across works by different authors in nineteenth-century Czech poetry. We show that the degree of recurrence is (1) comparable to Russian poetry, (2) noticeably weaker as compared to German and Spanish poetry and (3) strikingly weaker as compared to English poetry. Following the hypothesis formulated in Plecháč 2018, we offer a linguistic explanation for such findings: the size of rhyme repertory in a given language depends upon its inflection, the changes in word forms used to mark grammatical aspects such as voice, case, tense, mood, number, or gender. Roughly speaking—the more suffixes the language employs, the richer rhyme repertory gets as completely different words may rhyme only due to being followed by the same grammatical endings.

As one may expect, we find that in all the languages examined the tendency to share rhyme pairs is stronger between works that come roughly from the same period than between those distant in time. We thus try to represent each work as a vector defined by the relative frequencies of rhyme pairs and use machine learning techniques (Random Forest, Support Vector Machine) in order to classify them into time periods or poetic movements. Cross-validations of the models show that depending on what classes are being used, the accuracy varies between 0.6 to 0.9 and always outperforms the random baseline.The datasets come from:

- The Corpus of Czech Verse (Plecháč, P. - Kolár, R. 2015)
- The Gutenberg English Poetry Corpus (Jacobs 2018)
- Metricalizer corpus of German poetry (Bobenhausen-Hammerich 2015),
- Corpus of Spanish Golden-Age Sonnets (Navarro-Colorado 2017) & Diachronic Spanish Sonnet Corpus (Ruiz Fabo et al. 2017)
- Russian National Corpus (http://www.rusropora.ru)

References

Bobenhausen, K.-Hammerich, B. (2015). Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer. Langages 199, 67-87.

Jacobs, A. M. (2018). The Gutenberg English Poetry Corpus: Exemplary quantitative narrative analyses. Frontiers in Digital Humanities 5:5.

Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. Digital Scholarship in the Humanities 33(1), 112-127.

Plecháč, P.-Kolár, R. (2015). The Corpus of Czech Verse. Studia Metrica et Poetica 2(1), 107-118.

Plecháč, P. (2018). A collocation-driven method of discovering rhymes (in Czech, English, and French poetry). In M. Fidler - V. Cvrček (eds.), Taming the Corpus. From Inflection and Lexis to Interpretation. Cham: Springer, 79-95.

Reddy, S.-Knight, K. (2011). Unsupervised discovery of rhyme schemes. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. ACL, Portland, 77-82.

Ruiz Fabo, P.-Bermúdez Sabel, H.-Martínez Cantón, C. - Calvo Tello, J. (2017). Diachronic Spanish Sonnet Corpus (DISCO). Madrid, UNED. https://zenodo.org/record/1069844

Network analysis and Russian rhyme

David J. Birnbaum
Elise Thorsen

Conventions about the nature of phonetic similarity in Russian rhyme have varied over time. At the end of the explosively innovative "Silver age of Russian literature", Viktor Žirmunskij (1923) and Valerij Brjusov (1923) both recognized a prevailing tendency in poetic composition of the time to test the limits of rhyme, but with subtle differences: either as a shedding of an artificial apparatus to the bare essentials of rhyme or as an elaboration on the rule set defining the rhyme zone. Žirmunskij's and Brjusov's approaches to rhyme were formal, but not quantitative, nor, of course, computational. Mixail Gasparov (1984) later argued that rhyme is often contingent on outside conditions, including prosodic innovations of one era becoming marked archaisms to avoid in another. Gasparov's approach was explicitly quantitative, although also not computational. As in other dimensions of his quantitative studies of poetry, Gasparov produced a well-defined set of categories describing common deviations from exact phonetic repetition, and his pioneering work informs our computational exploration of imperfect end-rhyme.

Rhyme studies often operate at a corpus level, constructing large sets of rhyme pairs or nests that serve as training material for supervised machine-learning approaches to recognizing rhyme (see, for example, Plecháč 2018). This type of approach yields structured data that can be used for varied explorations of rhyme in verse, including network visualization, but a force-directed graph of word rhyme in a large corpus quickly becomes illegibly cluttered. Perhaps more importantly, where the object of study is the fluidity of rhyme over time, as is the case for us, an inventory of rhymes over a large corpus elides the instability and contingency of rhyming practice. Network graphs of word rhyme in individual poems (e.g., Lombardi 2013) are well suited to the visualization of rhyme, including imperfect rhyme, within a coherent domain, but because the number of rhyming words in a single poem is small, it is difficult to compare the network graphs of poems in a way that benefits meaningfully from network visualization.

Our approach to rhyme, and especially our focus on the fluid conventions of imperfect rhyme, relies on structural levels smaller than the word, such as phonetic segments and phonetic distinctive features (Birnbaum and Thorsen 2017, 2018, and 2019), and these close perspectives can be seen as complementing word-level approaches to rhyme. Segments and distinctive features lack lexical semantics, and therefore do not provide access to the types of meaning that are encoded or created by rhyme associations. But because the inventory of segments and distinctive features is (compared to lists of rhyming words) relatively small and relatively stable, they create a means of profiling rhyme that can be visualized with less clutter than network graphs of words, and that can be compared more easily. Our contribution to this panel, then, explores the opportunity for the graphic visualization of network rhyme relations below the word level and in the context of the mutability of rhyme conventions over the history of Russian verse.

Works cited

Birnbaum, David J. and Elise Thorsen. 2017. "Exploring inexact rhyme in Russian verse." Presented at "Plotting poetry | Machiner la poésie", 5-7 October 2017, Universität Basel.

Birnbaum, David J. and Elise Thorsen. 2018. "The automatic detection of Russian rhyme." Presented at "Plotting poetry 2: bringing deep learning to computational poetry analysis", 12-14 September 2018, Freie Universität Berlin.

Birnbaum, David J. and Elise Thorsen. 2019. "Rules-based and machine-learning approaches to identifying Russian rhyme." Presented at "Plotting poetry (and poetics) 3", 26-27 September 2019, ATILF, Nancy.

Brjusov, Valerij Jakovlevič. 1923. "O rifme." Review of: Viktor Maksimovič Žirmunskij, Rifma, ee istorija i teorija. Voprosy poètiki (Rossijskij institut istorii iskusstv), vyp. 2. Petrograd: Academia.

Gasparov, Mixail Leonovič. 1984. O?erk istorii russkogo stixa. Metrika, ritmika, rifma, strofika. Moscow: Nauka.

Lombardi, Thomas. 2013. "Network models of rhyme: 'The raven'." https://telombardi.wordpress.com/2013/08/08/network-models-of-rhyme-the-raven/

Plechá?, Petr. 2018. "A collocation-driven method of discovering rhymes (in Czech, English, and French poetry)." In M. Fidler, V. Cvrček, eds., Taming the corpus. From inflection and lexis to interpretation. Cham: Springer, 79-95.

Žirmunskij, Viktor Maksimovič. 1923. Rifma-ee istorija i teorija. Petrograd.

Rhyme network analysis in a non-canonical corpus of sonnets in Spanish

Pablo Ruiz and Helena Bermúdez

Rhyme constitutes a structural element for specific poetic forms, like sonnets. Given its relevance in versification, computational studies of rhyme can provide insight about style and intertextuality. We propose an exploration based on network analysis. Seeking representativity, we analyzed a corpus that draws from online anthologies, including non-canonical authors. We discuss challenges involved with this material and how networks can help analyze it.

Previous work employing network analysis of rhyme includes Houston (2017) for Victorian poetry, List (2016) for Old Chinese, Osinova (2016) for Russian, or Sonderegger (2011) for English poetry ca. 1900. Network nodes in these works are mainly rhyme-words, with edges indicating rhyme-pairs. Our contribution here is creating author networks, where rhyme-pairs are the edges between authors, which can be exploited to assess intertextuality between specific authors.

We analyzed the DISCO dataset of sonnets in

Spanish (Ruiz et al., 2018), which contains rhyme annotations, besides author metadata for birthplace and period. The corpus sources are online sonnet anthologies (García, 2005; 2006a; 2006b) which cover lesser-known authors besides some canonical ones, from Spain and Latin America, between the 15th and the 19th centuries. The corpus can be found at https://github.com/pruizf/disco/tree/ v3

Considerable rhyme-pair recurrence across authors was attested. We first discuss findings regarding central authors in the network, then some peripheral authors, and finally how peripheral patterns revealed textual errors in the corpus.

The most central author (Fig.1) is 16th-century Spanish author Juan de Arguijo, patron of a canonical author, Lope de Vega (Asensio, 1883: 26). Arguijo's centrality is not surprising; early scholars already noted that he had contemporary followers like Rodrigo Caro (Colon i Colon, 1841: 9). While not surprising, Arguijo's position can be seen as an indication of the network's sanity.

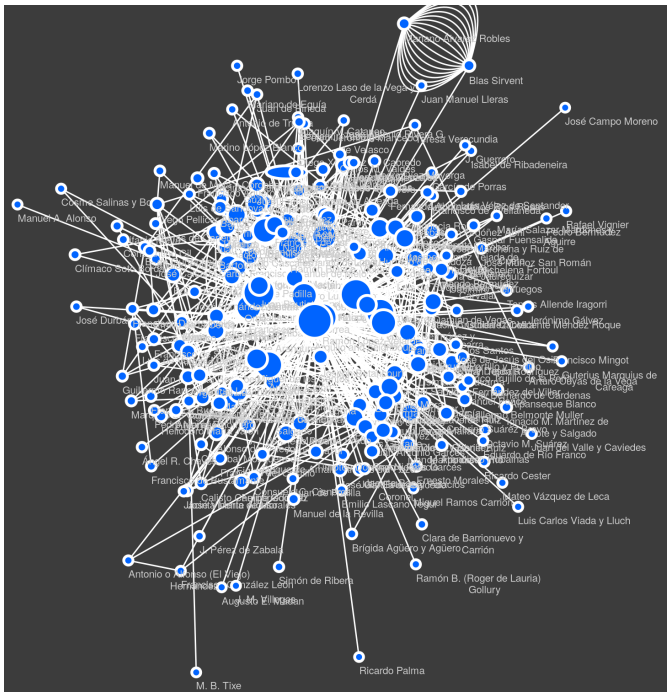


Figure 1. Most central author, Juan de Arguijo, and first neighbours grouped together (central node)

Peripheral areas of the network may also show interesting facts. In the top-right area of Fig. 1, two little-known authors strongly connected to each other appear. From discussions in López Cruces (2000: notes 95-96), we deduce that both wrote for a periodical that challenged authors to create sonnets using pre-established unlikely rhyme-words; the network seems to reflect this. To explore rhyme richness, a bipartite network with authors and rhyme-pairs as nodes was created. Non-central areas of this second network revealed data problems. Several disconnected components appeared (Fig. 3). These correspond to poems anthologized twice, but under two different author name variants The transmission of peripheral literature, via non-canonical sources, can pose "noise" problems. Finding these errors via networks highlights how they can help data curation: eccentricities led us to a close reading of some of the authors, then spotting textual problems. In summary, the paper shows how networks based on a heterogeneous, largely non-canonical corpus can confirm trends in it, besides showing interesting peripheral patterns and helping address textual problems inherent to the transmission of lesser-known literature. Future work includes integrating rhyme-data from a canonical sonnet corpus (Navarro-Colorado et al., 2016), for comparison. Also, evaluating rhyme variety per author considering the commonness of rhymes in different subnetworks, to assess originality vs. variety.

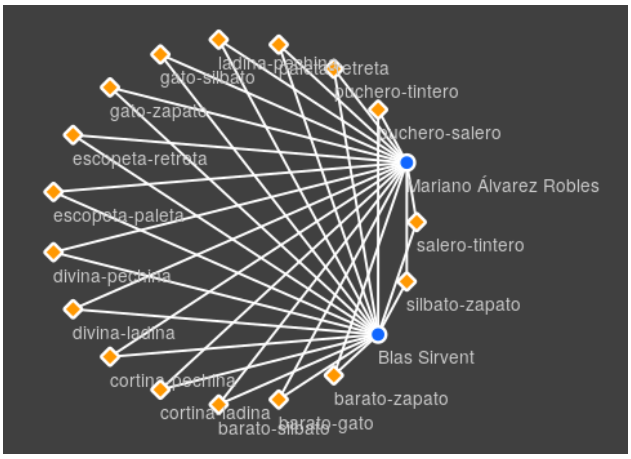


Figure 2. The bipartite network, filtered to isolate two strongly connected authors, also seen at the top-right of the author network on Fig. 1. Both authors (round nodes) share unusual rhymes (diamond nodes); from the literature we deduce they took part in a writing challenge involving those rhyme-pairs

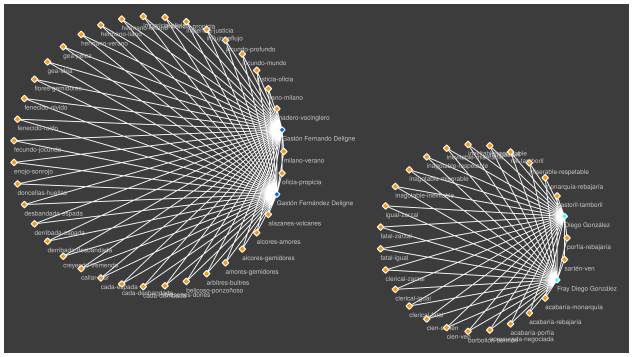


Figure 3. Disconnected components revealed textual problems in the dataset's sources

References

Asensio, J. M. (1883). Don Juan de Arguijo. Estudio biográfico. Madrid: Tipografía Guerrero. https://archive.org/details/AFol0681588/ (Accessed 13 Oct. 2019)

Colon i Colon, J. (1841). "Introducción". Sonetos de Don Juan de Arguijo. Sevilla: Francisco Álvarez. https://archive.org/details/AMont14417/ (Accessed 13 Oct. 2019)

García González, R. (2005). Sonetos del siglo XVIII. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xviii–0/html/. (Accessed 13 Oct. 2019)

García González, R. (2006a). Sonetos del siglo XV al XVII. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetosdel-siglo-xv-al-xvii–0/html/ (Accessed 13 Oct. 2019).

García González, R. (2006b). Sonetos del siglo XIX. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xix–0/html/ (Accessed 13 Oct. 2019).

Houston, N. (2017) Measuring Canonicity: a Network Analysis Approach to Poetry Anthologies. In Digital Humanities 2017. McGill University, Montreal Canada, pp. 476-478

List, J. (2016). Using Network Models to Analyze Old Chinese Rhyme Data, Bulletin of Chinese Linguistics, 9(2), 218-241. doi: https:// doi.org/10.1163/2405478X-00902004

López Cruces (ed.) (2000). Poesías jocosas, humorísticas y festivas del siglo XIX. Alicante: Biblioteca Virtual Miguel de Cervantes (Digital edition based on 1992 print edition from Alcodre, Alicante). http://www.cervantesvirtual.com/obra-visor/poesias-jocosas-humoristicas-y-festivas-del-siglo-xix–0/html/ (Accessed 13 Oct. 2019)

Navarro-Colorado, Borja, María Ribes Lafoz and Noelia Sánchez (2016): Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation, in Proceedings of the Language Resources and Evaluation Conference. http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf (Accessed 14 Oct. 2019)

Ruiz, P., Bermúdez, H., Martínez, C. I., González-Blanco, E., & Navarro-Colorado, B. (2018). The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings. In Digital Humanities 2018: Conference Abstracts. Red de Humanidades Digitales, Mexico City, pp. 486-489. http://e-spacio.uned.es/fez/eserv/bibliuned:363-Pruiz3/Ruiz_Fabo_Pablo_DISCO.pdf (Accessed 14 Oct. 2019)

Sozinova, O. (2016). Complex Networks-Based Approach to Russian Rhyme History Description: Linguostatistics and Database: In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 891-893.

Sonderegger, M. (2011). Applications of graph theory to an English rhyming corpus, Computer Speech & Language, Volume 25, Issue 3, 2011, pages 655-678, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2010.05.005.

Linked Art: Networking Digital Collections and Scholarship

Kevin Page

kevin.page@oerc.ox.ac.uk
Oxford e-Research Centre, University of Oxford, United Kingdom

Emmanuelle Delmas-Glass

emmanuelle.delmas-glass@yale.edu
The Yale Center for British Art, Yale University, USA

David Beaudet

d-beaudet@nga.gov
The National Gallery of Art, Washington D.C., USA

Samantha Norling

snorling@discovernewfields.org
Indianapolis Museum of Art at Newfields, Indianapolis, USA

Lynn Rother

lynn.rother@leuphana.de
Museum of Modern Art, New York, USA Leuphana Universität, Germany

Thomas Hänsli

thomas.haensli@uzh.ch
University of Zurich, Switzerland ETH Zurich, Switzerland

Keywords: Linked Data, Art, Art History, Art Provenance, Cultural Heritage Institutions
Topics: Global, English, 18th Century, 19th Century, 20th Century, data modeling, linked (open) data, Art history, Humanities computing

Panel Overview: Linked Art - Networking Digital Collections and Scholarship

Linked Art[1] is a major new initiative by which art museums will publish information about their collections as interconnected open data. Building upon long standing interdisciplinary thinking from the digital humanities and information engineering, Linked Art is an international collaboration across twenty-four institutions identifying focussed, practical models which meet their requirements on a sustainable basis.

The central aim of Linked Art is the development and application of Linked Data to cultural heritage collections, with an emphasis on works of art and their provenance. Linked Data will provide the foundation for multi-modal digital scholarship across these rich collections; as an open data standard, Linked Art provides consistent, structured ways for arts institutions to publish art-related data where, in many cases, there has not been a consistent shared model to date.

This panel presents a range of perspectives representative of the collaborative intersections found in the Linked Art community: with speakers from universities and art museums; who are practitioners and academics; on topics ranging from implementation, to curation, and research.

The panel will takes the form of six position papers, outlined below, followed by questions and answers between the audience and panel. In doing so, Linked Art seeks to engage with the Digital Humanities community, building capacity for future collaborative implementations and research investigations.

1. Linked Data and Open Data in Cultural Heritage

Emmanuelle Delmas-Glass, The Yale Center for British Art, Yale University, USA.

Cultural heritage institutions have a great deal to gain from deeply engaging in the networked environment. They have poured many resources in the digitization of their collections for the benefit of their audiences, from students to experts, who want to have access to more online material of a higher quality. The current landscape of cultural heritage knowledge on the Web, however, is very much siloed, which both harms the relevance of individual institutions as well as the overall state of scholarship which might use that knowledge.

This paper reflects on the main challenges that cultural heritage institutions face when it comes to publishing their collections descriptions as Linked Open Data resources. Some challenges can be due to friction between a declared digital mission and the resources allocated, which might seem to be in competition with other institutional priorities. Other memory institutions are still in the process of understanding that managing their knowledge and data - so it can then be leveraged for the Semantic Web - needs to be a core data curation activity. It is also partly due to the lack of entry-level technology and ontology resources that has prevented museums from engaging more deeply with Linked Open Data.

This talk will give an overview of previous initiatives and technologies intended to open up access to cultural heritage institutions, particularly art museums, including the International Council of Museums Committee for Documentation Conceptual Reference Model (CIDOC CRM, also an ISO standard); The American Art Collaborative (AAC); PHAROS, the International Consortium of Photo Archives; the Art and Architecture Thesaurus (AAT); the Union List of Artist Names (ULAN); and the International Image Interoperability Framework (IIIF).

In the context of the successes and limitations of these efforts, this paper will outline the strategy taken by Linked Art, which is both a standard-based data model and a community, which has emerged from earlier work at the Getty Research Institute. This presentation will challenge the traditional paradigm which has large and wealthy institutions succeed in the face of structural challenges. In this era of hyper connectedness, the solution to museums' relevance in the Web cannot be developed by a lone institution, and indeed the model that Linked Art promotes is instead based on inclusion with the goal to create institutional and individual partnerships. The other precept that the Linked Art data model advocates for is usability over absolute data completeness, and this talk will go over some specific data modeling principles that allow balance between the requirements of the institution, domain knowledge experts, technologists who will implement the standard, and scholars and other users of Linked Art.

Emmanuelle Delmas-Glass is the Collections Data Manager at the Yale Center for British Art, a member of the IIIF Operating Committee, and co-chair of the Linked Art working group of the International Council of Museums' Committee for Documentation (ICOM CIDOC). She oversees the creation, access to and distribution of the museum's collections information and metadata, playing the lead role in ensuring the intellectual and technical integrity of the collections data and metadata.

2. Conceptual models meet practice and scholarship: conventions, standards, and technology

Kevin Page, Oxford e-Research Centre, University of Oxford, UK.

The utility of a data model is dependent on its usage, and the context in which this use occurs. While conceptual models play an important role in providing a framework within which different data sets can be

consistently represented and combined, in practice the resulting information structures can be perceived as complex or unwieldy, which can stifle adoption in cultural heritage institutions.

Specialisms in scholarship and practice bring differing requirements and benefits to the structuring of collections data and its analysis, which are similarly defined by use - the operational needs of a collection or library are different from that of an exhibition, or of the intellectual needs of academic study. The challenge, then, is in respecting and encouraging these different 'information perspectives', whilst benefiting from their intersections where they occur, and managing the complexity of the systems and organisational implications. In addition, we recognise scholarship within cataloguing and curation activities, whilst appreciating these have different - but complementary - information needs and outputs from academic study of the collections. Digital tools and methods should reflect these activities, roles and specialisms, rather than constrain them. Doing so can ease adoption of digital approaches alongside existing established practice, increasing the sources of compatible structured information, and achieving overall progress through the combination of multiple information intersections.

In considering the above, this paper reflects upon more nuanced notions of authority, standards, and how these are realised technologically; moving from necessary 'on the wire' interoperability to a progression from local practice, through emergent community conventions, to international standards bodies. It can be beneficial for different stages of maturity to exist simultaneously across distinct but complementary information structures, reflecting the communities of specialist practice and scholarship who are using the data. The technologies of Linked Open Data, including RDF and ontologies, provide a flexible foundation through which we can realise such an iterative and incremental approach to standardisation, and in which alternative information perspectives can co-exist.

Linked Art recognises a further first-class perspective: that of the software developer writing (potentially for, or with scholars) applications which consume collection data provided by cultural heritage institutions. It adopts the principles of Linked Open Usable Data, as proposed by Rob Sanderson of the Getty Research Institute, to create a profile of the CIDOC CRM tailored to this information perspective, and in which established practice - for example, in the use of AAT - can be respected.

Kevin Page is Associate Faculty at the University of Oxford e-Research Centre. He was Technical Director of the Oxford Linked Open Data (OXLOD) project, a prototype for the use of CIDOC CRM across the Gardens, Libraries, and Museums of the University of Oxford. Kevin is Principal Investigator of the Linked Art Research Network and Linked Art II project, both funded by the UK's Arts and Humanities Research Council, and a member of the Linked Art Editorial Board.

3. Practicing Linked Art: Evolving Art Data at the National Gallery of Art

David Beaudet, National Gallery of Art, Washington D.C., USA.

The National Gallery of Art (NGA) seeks to serve the United States in a national role by preserving, collecting, exhibiting, and fostering the understanding of works of art, at the highest possible museum and scholarly standards. As such, the institution is frequently engaged in activities that require production of the structured data describing our collection, the associated media, and writings related to the collection's historical significance. These demands for data manifest in a variety of ways. Images are requested en masse, data sets are requested in support of research, and both data and media are increasingly requested in support of on-site and remote visitor experiences as well as for analytical purposes.

In recognition of the need to automate the accurate publication of data about art, artists, depictions of art, and associated media, the NGA's department of Analytics and Enterprise Architecture has been seeking data modeling and dissemination standards that are accepted by the cultural heritage art community in order to ensure the greatest reach possible from its automated art data services. The NGA has selected Linked Art as a strategic data standard.

Standardizing formats and exchanges of data through automated means has paid dividends for many industries in the past and continues to do so. For example, the IIIF standards[2] for image sharing have expanded the reach of deep zoom technologies and image collections across the cultural heritage sector and it is expanding into other communities[3]. Whilst data standards are not in themselves novel, no prevailing broadly adopted standard for modeling art data exists. Linked Art seeks to fill that gap.

As part of its participation in the Linked Art community, the NGA is evolving an existing art data interface, one that currently provides collection data and images to its Conservation Space system, to use the Linked Art standards as a proof of concept. This paper will give details of that implementation, alongside the basics of the Linked Art model[4] - for representing artworks, people, and depiction. The evolving, open source[5], public-facing NGA interface already provides art data to a location-aware mobile app[6] available for

visitors to download on their iOS devices, which will be demonstrated.

A solution architect with the National Gallery of Art since 2005, David Beaudet designs and builds technical solutions for authoring and publishing rich art imagery, content, and metadata. David is a member of the editorial board of Linked Art and collaborates on the IIIF Discovery API.

4. The Linked Art of Georgia O'Keeffe: Collections Across Institutional Boundaries

Samantha Norling, Indianapolis Museum of Art at Newfields, Indianapolis, USA.

The creation and publication of linked open data allows for previously siloed data to be connected with other related data on the Web, breaking down institutional barriers and facilitating research and new scholarship - in traditional and digital formats - that may not have previously been possible. With members and community participants representing over 20 different arts-related organizations, the Editorial Board for the developing Linked Art data model for describing art collections reflects this barrier-breaking nature of linked data.

In order to showcase the connections that can be made when multiple institutions publish their collections data in a consistent format and utilize shared vocabularies, members of the Linked Art community collaborated to create a cross-institutional sample data set. The artist Georgia O'Keeffe was selected to serve as the common thread for the data to be contributed by participating institutions. With an emphasis on relationships, the linked data collected for the showcase naturally expanded to include not just O'Keeffe and her artworks, but also the works of her contemporaries, the exhibitions in which the artworks were exhibited, and the various organizations and individuals that had participated in provenance events in the lifecycle of the artworks.

The development of the Linked Art O'Keeffe data set serendipitously coincided with the launch of the Georgia O'Keeffe Museum's (GOKM) beta version of the GOKM's Collections Online, which was built on a linked data foundation. The GOKM expressed the importance of linked data to their digital strategy to make it possible for "meaningful connections to be expressed across different types of collections (artworks, archival items, books, etc.) to establish a more complete understanding of Georgia O'Keeffe's life, work, and contexts."

Drawing on both the GOKM's Collections Online[7] and the co-constructed O'Keeffe showcase data set that is now available publicly within the Linked Art GitHub repository[8], this paper and panel presentation will explore the network of Linked Art data with Georgia O'Keeffe at the center. Following intersections between and connections across data sets, the exploration will follow a path through the many relationships identified that link artworks, archives, exhibitions, and people within O'Keefe's linked data network. In discussing some of the specific relationships identified between institutional collections, the exploration will also highlight key patterns within the Linked Art data model. The Georgia O'Keeffe showcase data set, while small in scale, demonstrates the potential for the Linked Art data model to facilitate the creation of new connections between collections of all types - connections that cross institutional boundaries and facilitate scholarship at the intersections between GLAM collections.

As Digital Collections Manager at the Indianapolis Museum of Art at Newfields, Samantha Norling manages digital assets and data related to the museum's art, archival, and horticultural collections. A trained librarian and professional archivist working within an art museum, Samantha is particularly interested in digital projects that break down barriers between GLAM institutions. She is a member of the Linked Art Editorial Board.

5. The History of Art is Linked but the Data Is Not: Georgia O'Keeffe, Provenance and Scholarship

Lynn Rother, Leuphana Universität, Germany.

The history of artworks is linked. They were produced by the same artists, traded by the same dealers, collected by the same people, transferred, looted or confiscated by the same entities while eventually finding their permanent home in the same museums - or not. To date, these links across museum collections are only visible to the few scholars or experts studying the artworks' history of ownership. But the field of provenance research has matured enough to enable and support structuring and aggregating provenance records as Linked Data.

Though shaped by complex and diverse contexts, an artwork's provenance record can be broken down into empirical data consisting of objects, protagonists, dates, locations and types of transactions. To this day, however, the majority of museums record the valuable information harvested through time-consuming and resource-intensive provenance research within their collection management systems without machine-readable structure, hindering the analysis and linking of the data across institutions on a larger scale. Digital humanities tools offer the potential to standardize, aggregate, and consider the museum accumulated provenance data broadly to reveal new stories about the global circulation and displacement of artworks and nuance the existing histories of collecting and art market practices.

As museum objects and their movements through time and space tell stories beyond object-based art historical research and collection cataloguing, this paper will elaborate on the potential of Linked Art for provenance research and for scholarship in related fields. The intertwined histories of selected works by the American Modern artist Georgia O'Keeffe - from different museum collections including the Georgia O'Keeffe Museum in Santa Fe, New Mexico and The Museum of Modern Art (MoMA) in New York - will serve as an example.

In particular, MoMA's acquisition and deaccession of O'Keeffe's Kachinas - representations of Pueblo and Hopi spirits used in ceremonies and rituals and therefore considered culturally sensitive objects in museum collections - will show how structured provenance data of museums using Linked Art can benefit related research fields such as the histories of collecting and art market practices but also museum, Native American, and Indigenous studies.

Lynn Rother is the Lichtenberg-Professor for Provenance Studies at Leuphana Universität, Lüneburg, Germany, and a member of the Linked Art Editorial Board. Previously, Lynn was Senior Provenance Specialist at The Museum of Modern Art, New York, where she oversaw provenance research, procedures, documentation, digital strategies and funding in conjunction with all curatorial departments regarding works in the Collection, loans, acquisitions, and deaccessions.

6. Topographia Helvetiae: Linked Art in Switzerland

Thomas Hänsli, University of Zurich / ETH Zurich, Switzerland.

The historical view of Switzerland's nature has been defined by artworks describing a 'visual topography' of the Alpine country long before the emergence of a broader touristic interest for Switzerland across Europe. Paintings, drawings, prints, and photographs depicting alpine landscapes, natural monuments, picturesque villages and much more shaped the perception of Swiss landscapes both nationwide and beyond.

The Swiss Art Research Infrastructure (SARI) provides unified and mutual access to Swiss collection data, research data, and digitised visual resources from museums, archives, collections, as well as academic and public research institutes, based on a Linked Open Data network. Being part of a national research infrastructure programme, its mission is to combine the unique scholarly expertise from specialised research institutions beyond technical, linguistic, and institutional borders and to enhance the visibility and accessibility of Switzerland's valuable collections and research resources.

The aggregation and access of these visual resources is fundamental for browsing and understanding the evolution of the framing of Switzerland. The project »Bilder der Schweiz« (Views of Switzerland), developed under the aegis of SARI, provides a unique access point to topographic artworks and photographs from the eighteenth to the early twentieth century from major Swiss libraries, museums, and private collections. Starting from mass-printed, but hand-coloured vedute of the so-called 'Schweizer Kleinmeister' -- an affordable visual medium to propagate a canonised view of Switzerland that gained increased popularity over time -- the research portal includes related materials such as landscape paintings, drawings, printed textual sources, travel guides, travel journals, and further materials related to artist production, printing and marketing of printed 'vedute'.

The project will provide unified access to these heterogeneous materials from comparable, but technically different, institutional repositories following the Linked Art data model; and in doing so enhance public visibility of these little known, but widely consumed 'Schweizer Kleinmeister', and make them accessible to scholars.

The project also provides an excellent test of the Linked Art framework's flexibility, assessing the model's ability to describe specific non-mainstream subject-based collections. This paper will reflect upon the overall transformation of a diverse selection of data sources into a Linked Art compliant format; with a specific focus on the advantages and drawbacks of the framework when describing tight semantic integration between the expression (the depicted visual apparatus) and the content (the perspective over the object). Finally, a reflection over the possible coexistence of multiple levels of description of an object, each level addressing specific communities, will be presented.

Head of gta Digital (ETH Zurich, 2011) and director of the Swiss Art Research Infrastructure (University of Zurich, 2017), Thomas Hänsli has authored the strategy for a national research network and co-authored the development and implementation of comprehensive reference data models based on CIDOC-CRM. He is responsible for the implementation of several Linked Open Data projects in Switzerland and a member of the Linked Art Editorial Board.

[1] https://linked.art/
[2] https://iiif.io
[3] https://iiif.io/community/groups/
[4] https://linked.art/model/
[5] https://github.com/NationalGalleryOfArt/data Services

[6] https://apps.apple.com/us/app/national-gallery-of-art-dc/id1455655720
[7] https://collections.okeeffemuseum.org
[8] https://github.com/linked-art/showcase1/

Scribes, Scholars, & Scripts: creating a Digital Humanities community through crowdsourcing

Emily Esten
estenemily@gmail.com
University of Pennsylvania Libraries, United States of America

Samantha Blickhan
samantha@zooniverse.org
Adler Planetarium

Will Noel
wgnoel@pobox.upenn.edu
University of Pennsylvania Libraries, United States of America

Marina Rustow
mrustow@princeton.edu
Princeton University

Keywords: open data, crowdsourcing, multilingual DH, text analysis, project management
Topics: Global, English, 5th-14th Century, 15th-17th Century, 18th Century, crowdsourcing, public humanities collaborations and methods, History, Library & information science

How do you organize a research project around manuscript fragments? Digital collections can allow teams to create a shared online space in which images can be hosted, but the types of research questions that individual scholars may want to ask can be as fragmented as the objects of interest themselves. In this panel, we will discuss Scribes of the Cairo Geniza (https://www.scribesofthecairogeniza.org), a collaboration between the University of Pennsylvania Libraries and the Zooniverse (https://www.zooniverse.org), the world's largest platform for online crowdsourced research. The project invites the public to help classify and transcribe fragments from the Cairo Geniza, a corpus of 350,000 fragments primarily from the 10th-13th centuries, found in a storeroom (or 'geniza') of the Ben Ezra synagogue in Fustat. The panelists will discuss the process of designing, developing, and running a large, multi-institution online crowdsourcing project from the following perspectives: project manager, web developer, content specialist, and data specialist.

Replication and Computational Literary Studies

Christof Schöch
schoech@uni-trier.de
University of Trier, Germany

Karina van Dalen-Oskam
karina.van.dalen@huygens.knaw.nl
Huyghens ING and University of Amsterdam, The Netherlands

Maria Antoniak
maa343@cornell.edu
Cornell University, USA

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

David Mimno
Cornell University, USA

Keywords: replication, reproduction, computational literary studies
Topics: Global, English, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), text mining and analysis, Literary studies

Technical Note

A PDF version of this contribution is available at Humanities Commons: https://hcommons.org/deposits/item/hc:30439 Several versions of this contribution in structured formats, including figures and bibliographical

data, are available from Zenodo: http://doi.org/10.5281/zenodo.3893428

Panel Overview

The "replication crisis" that has been raging in fields like Psychology (Open Science Collaboration 2015) or Medicine (Ioannidis 2005) for years has recently reached the field of Artificial Intelligence (Barber 2019). One of the key conferences in the field, NeurIPS, has reacted by appointing 'reproducibility chairs' in their organizing committee[1]. In the Digital Humanities, and particularly in Computational Literary Studies (CLS), there is an increasing awareness of the crucial role played by replication in evidence-based research. Relevant disciplinary developments include the increased importance of evaluation in text analysis and the increased interest in making research transparent through publicly accessible data and code (open source, open data). Specific impulses include Geoffrey Rockwell and Stéfan Sinclair's re-enactments of pre-digital studies (Sinclair and Rockwell 2015) or the recent replication study by Nan Z. Da (Da 2019). The paper has been met by an avalanche of responses that pushed back several of its key claims, including its rather sweeping condemnation of the replicated papers. However, an important point got buried in the process: that replication is indeed a valuable goal and practice.[2] As stated in the Open Science Collaboration paper: "Replication can increase certainty when findings are reproduced and promote innovation when they are not" (Open Science Collaboration 2015, 943).

As a consequence, the panel aims to raise a number of issues regarding the place, types, challenges and affordances, both on a practical and on a policy or community level, of replication in CLS. Several impulse papers will address key aspects of the issue: recent experience with attempts at replication of specific papers; policies dealing with replication in fields with more experience in the issue; conceptual and terminological clarification with regard to replication studies; and proposals for a way forward with replication as a community task or a policy issue.

Contribution 1: "A typology of replication studies", by Christof Schöch

This contribution aims to provide orientation about the range of existing replication studies, based on a simple typology. The typology describes the relationship between an earlier study and its replication in terms of four key variables: the research question, the method of analysis (including the implementation of that method) and the dataset used.[3] For each of these variables, a replication study can attempt to operate either in the same way as the previous study, or in a different way. Note that the typology is not meant to establish these distinctions in a purely binary fashion: rather, as data or methods are never entirely identical or completely different from the earlier study, the extreme points in the typology are meant to open up a gradient of pratices. In addition (and this aspect might be unique to the Digital Humanities), replication studies can be described as involving crossing the boundary between non-digital and digital research or between qualitative and quantitative research.

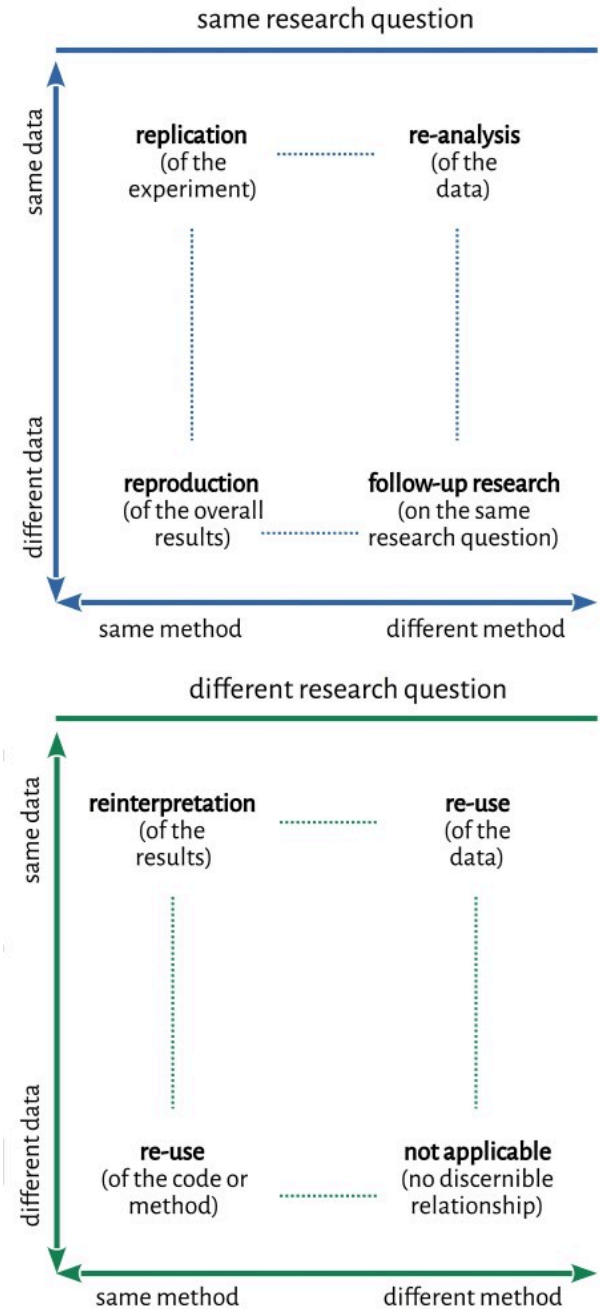


Figure 1: Typology of repeating research

At the most fundamental level, such a typology structures the field and provides a clearly-defined terminology and systematic relations between the various types. For example, replication vs. reproduction or re-analysis vs. follow-up research. Such a shared understanding is useful because each type of replication study comes with its own objectives, requirements and challenges as well as their own place and function in the research process. A replication study strictly repeating key aspects of an earlier study will be most useful in reviewing and quality assessment, while follow-up research more loosely modeled after an earlier study may instead have important methodological implications or lead to knew domain knowledge. This is true despite the fact that in reality, there is more to consider than a few binary categories when describing a given replication study.

But understanding replication through such a typology can have an impact on the field of DH in a number of additional ways. It can provide guidance when publishing research and help clarify what needs to be provided (in terms of data, code and contextualization in prose) in order for a study to be amenable to a specific type of replication. It can help assess the merits and limitations of a given replication study to assess whether, given the stated objectives of the authors, they have employed a suitable type of replication strategy. And it can support designing a replication study and clarify what data, code and contextual information needs to be obtained or reconstructed in order to perform a specific type of replication.[4]

Beyond this, such a typology can contribute to better define the relationship between replication in the strict sense and related efforts like benchmarking and evaluation studies. Finally, because such a typology makes it easier to identify similar studies across disciplinary boundaries, it may help us as a field learn more quickly from other fields with a longer tradition in (specific types of) replication studies. In this way, such a typology of replication studies can contribute to establishing replication as a well-understood part of Computational Literary Studies.

Contribution 2: "Replication to the Rescue: Funding Strategies", by Karina Van Dalen-Oskam

The Dutch Research Council (NWO) is the first funding agency to take the initiative for a pilot programme for Replication Studies. Their aim is "to encourage researchers to carry out replication research. NWO wants to gain experience that can lead to insights into how replication research can be effectively included in research programmes. That experience should also lead to insights into and a reflection on the requirements that NWO sets for research in terms of methodology and transparency."[5] The first two rounds of funding were in 2017 and 2018, and were aimed at the Social Sciences and Medical Sciences. In the third round in 2019, the Humanities were included.

This was done after a heated discussion in Nature between Rik Peels and Lex Bouter (chair of the Replication Studies Programme Committee) on the one hand, and Bart Pender, Sarah de Rijcke and J. Britt Holbrook on the other. Peels and Bouter (2018) started of with a note titled "Humanities need a replication drive too". De Rijke and Penders (2018), both scholars from Science and Technology Studies, countered with the call to "Resist calls for replicability in the humanities". They argue that quality criteria are crucially different in the humanities and the sciences.[6]

NWO went ahead with including the humanities in the call for replication studies, stating they are aware that not all humanities research is suitable for replication. NWO "expresses no preference or opinion about the value of various methods of research. Where possible it wants to encourage and facilitate the replication of humanities research: this should certainly be possible in the empirical humanities."[7] In March 2020, seven proposals were awarded funding, but none of these can be called typical Humanities projects.[8] How many submissions were received from humanities applicants - did scholars indeed resist, as Pender and De Rijcke advised? And in a wider context: How do Dutch Humanities scholars evaluate the new possibility? And does this agree with the reception in the growing and very active Dutch Digital Humanities community?

In my short impulse paper, I will reflect on what we can learn from the explicit invitation to the humanties to apply for funding for replication studies. What does this tell us about the status of humanities research in the Netherlands, and more specifically about the role of the Digital Humanities? I will pay special attention to the opportunities these developments may have for Computational Literary Studies. Should we consider the situation as "Funding Strategies to the Rescue: Replication", so a turning around of the title of my talk?

Contribution 3: "Replication of quantitative and qualitative research - a case study", by Fotis Jannidis

Literary studies always had an empirical side - 'empirical' in the broader sense, that claims and counterclaims are substantiated by referring to specific parts of texts. These text segments are regarded as indicators which in their sum make a more general point plausible, for example the use of specific terms to validate a hypothesis about a text. Therefore, the concept of replication warrants a wider understanding in Computational Literary Studies. The prototypical center is the quantitative replication of quantitative

research, but it also includes quantitative replication of qualitative philological research: Using the same indicators to validate the same hypothesis but moving the research into an empirical framework. Seen in the context of the discussion of mixed methods, this is a specific case of 'triangulation'. Triangulation refers to "the application of different data analysis methods, different data sets, or different researchers' perspective to examine the same research question or theme" (Bergin 2018, 29). But here, data sets and data analysis methods overlap strongly, while the research framework is changed from hermeneutic to quantitative.

Our case study is an attempt to replicate research on the complexity of language in German dime novels, published by Peter Nusser (Nusser 1981), and it demands both kinds of replication. Nusser describes the language of dime novels on three levels: vocabulary, syntax, and phrases. The work on vocabulary and syntax is quantitative, while the analysis of phrases is qualitative. The replication of the quantitative parts is made more difficult by the fact that the results which Nusser reports have actually been produced by another author in the context of an unprinted exam thesis which seems to be lost for the moment. So a lot of information is missing, and we can only make educated guesses: the exact corpus design (for high literature only the authors are given and for dime novels only the series), the strategies of tokenization and sentence splitting, the exact formula for calculating specific values, etc. The qualitative research is enumerating many phrases which are seen as examples of clichés and there is no explicit comparison with high literature. So a quantification must try to operationalize the concept of cliché and then compare retrieval results between dime novels and high literature.

As is well known in Computational Literary Studies, operationalization as an instance of formal modeling usually covers some aspects that are part of the intuitive notion, while others are excluded for the time being and it is one goal to reduce the loss (Moretti (2013); for a counterposition see Underwood (2019, 181)). In a replication, the loss may be responsible for the difference in outcome. In view of all these difficulties it could seem an unnecessary endeavor to replicate the research, but Nusser's study had a huge influence on the assessment and evaluation of popular literature in German studies for almost four decades.

Contribution 4: "Reliable methods for text analysis", by Maria Antoniak and David Mimno

If we are to make reproducible computational claims about literary texts, we need methods that lend themselves to robustness and reliability. Here we focus on the case study of word embeddings, which analyze collections of documents and produce numeric representations of words. Although these methods are powerful, they are also at high risk for problems with reproducibility: they are complicated enough to be essentially "black boxes", yet they are also known to be highly sensitive to text curation choices, parameter settings, and even random initializations (Antoniak and Mimno 2018). How can we assure researchers and their audiences that seemingly small changes would not alter or even reverse their findings?

Embedding vectors are useful for their ability to operationalize thick cultural concepts. For example, the resulting vectors have been used to measure shifts in word meaning over time and geographic areas (e.g. Hamilton, Lescovec, and Jurafsky (2016); Kulkarni, Perozzi, and Skiena (2016)). Several studies have shown that embeddings can encode gender biases by probing embedding spaces using carefully chosen seed words (Gordon and Van Durme (2013); (Bolukbasi et al. 2016a); Caliskan Islam, Bryson, and Narayanan (2016)). Subsequent work in natural language processing has focused on removing biases from an embedding model (Bolukbasi et al. (2016b); Sutton, Lansdall, and Cristianini (2018)). In this context, the concern is the downstream impact of bias on systems that use embeddings, but similar work can also be motivated from an upstream perspective, as a means of studying bias in collections.

Researchers from the humanities and social sciences use embeddings to provide quantitative answers to otherwise elusive political and social questions about the training corpus and its authors (e.g. Kozlowski, Taddy, and Evans (2019)). These bias detection techniques were originally intended to measure the bias encoded in a trained embedding; they were not originally tested to measure the bias of a corpus and make comparisons between corpora.

We probe the stability of these measurements by testing two popular bias detection methods (Bolukbasi et al. (2016a); Caliskan Islam, Bryson, and Narayanan (2016)) on sets of automatically constructed seed sets. These sets were constructed by randomly selecting a target term and then including its N nearest neighbors in the set; this process more closely approximates a real seed set, constructed by a scholar interested in a particular concept, than a random set of seeds. We find that bias detection techniques via word embeddings are susceptible to variability in the seed terms, in both their order (alternative pairings of seeds from two sets can significantly change the ability of the method to capture a single bias subspace) and semantic similarity (the more similar seeds set are to each other, the more difficult it is to measure their biases). If done carefully, bias detection using embeddings is feasible

even for small, subdivided collections and can provide a promising tool for differential content analysis, but we encourage error analysis of the seed terms.

We further highlight a central inconsistency in these bias detection methods. While these methods seek to measure biases in datasets, the researcher-selected seeds themselves can contain a variety of biases. For example, the seeds used for racial categories often include lists of names that are "African American" or "European." Such lists can be both reductive and essentializing. In addition, some seed sets contain confounding terms, e.g., contain a gendered term in a seed set for "domestic work" that is then used to measure gender bias. If the seed set for "domestic work" appears closer to the gender that it contains, it will be impossible to say whether that bias exists because of the training corpus or because of the inclusion of the gendered seed.

This case study highlights the reversal in perspectives when techniques from natural language processing and machine learning are re-purposed for studies of specialized datasets. Some working assumptions from the machine learning community (e.g. large size of training set) are broken in the humanities context, where datasets are non-expandable and are the primary focus of the study, rather than a generalized training set for downstream applications. The stability and robustness of these repurposings should not be assumed but rather should be reanalyzed for the particular new contexts.

Bibliography

Antoniak, Maria, and David Mimno. 2018. "Evaluating the Stability of Embedding-Based Word Similarities." Transactions of the Association for Computational Linguistics 6. https://transacl.org/ojs/index.php/tacl/article/view/1202.

Barber, Gregory. 2019. "Artificial Intelligence Confronts a 'Reproducibility' Crisis." Wired, 2019. https://www.wired.com/story/artificial-intelligence-confronts-reproducibility-crisis/.

Bergin, Tiffany. 2018. An Introduction to Data Analysis. Quantitative, Qualitative and Mixed Methods. London: Sage.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016a. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." In NIPS. https://arxiv.org/abs/1607.06520.

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016b. "Quantifying and Reducing Stereotypes in Word Embeddings." In Data4Good: Machine Learning in Social Good Applications, edited by Kush R. Varshney. http://arxiv.org/abs/1606.06121.

Caliskan Islam, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2016. "Semantics Derived Automatically from Language Corpora Necessarily Contain Human Biases." In CoRR. http://arxiv.org/abs/1608.07187.

Da, Nan Z. 2019. "The Computational Case Against Computational Literary Studies." Critical Inquiry 45 (3): 601-39. https://www.journals.uchicago.edu/doi/abs/10.1086/702594.

De Rijke, Sarah, and Bart Penders. 2018. "Resist Calls for Replicability in the Humanities." Nature 560 (29). https://www.nature.com/articles/d41586-018-05845-z.

Deckker, Harrison, and Paula Lackie. 2016. "Technical Data Skills for Reproducible Research." Edited by Lynda M. Kellam and Kristi Thompson. Databrarianship: The Academic Data Librarian in Theory and Practice. https://escholarship.org/uc/item/8qb2q8fk.

Gómez, Omar S., Natalia Juristo, and Sira Vegas. 2010. "Replication, Reproduction and Re-Analysis: Three Ways for Verifying Experimental Findings." In RESER '2010 Cape Town.

Gordon, Jonathan, and Benjamin Van Durme. 2013. "Reporting Bias and Knowledge Acquisition." In Proceedings of the 2013 Workshop on Automated Knowledge Base Construction. https://doi.org/https://doi.org/10.1145/2509558.2509563.

Hamilton, William L., Jure Lescovec, and Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Edited by Hamilton, Jure Lescovec, and Dan Jurafsky. CoRR. http://arxiv.org/abs/1605.09096.

Hüffmeier, Joachim, Jens Mazei, and Thomas Schulte. 2015. "Reconceptualizing Replication as a Sequence of Different Studies: A Replication Typology." Journal of Experimental Social Psychology, 81-92. https://www.sciencedirect.com/science/article/pii/S0022103115001195.

Ioannidis, J.A. 2005. "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research." JAMA, no. 294/2: 218-28. https://doi.org/https://doi:10.1001/jama.294.2.218.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. "The Geometry of Culture: Analyzing Meaning through Word Embeddings." American Sociological Review 84 (5). https://doi.org/https://doi.org/10.1177%2F0003122419877135.

Kulkarni, Vivek, Bryan Perozzi, and Steven Skiena. 2016. "Freshman or Fresher? Quantifying the Geographic Variation of Internet Language." In Proceedings of the Tenth International Conference on Web and Social Media. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13121.

Moretti, Franco. 2013. "'Operationalizing': Or, The Function of Measurement in Modern Literary

Theory." Stanford Literary Lab Pamphlet 6. https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf.

Nusser, Peter. 1981. Romane für die Unterschicht. Der Groschenroman und seine Leser [1973]. Stuttgart: Metzler.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." Science, no. 349 (6251). https://science.sciencemag.org/content/349/6251/aac4716.

Peels, Rik, and Lex Bouter. 2018. "Humanities Need a Replication Drive Too." Nature 558 (372). https://www.nature.com/articles/d41586-018-05454-w.

Peels, Rik, and Lex Bouter. 2018a. "Replication Is Both Possible and Desirable in the Humanities, Just as It Is in the Sciences." LSE Impact Blog, 2018. http://blogs.lse.ac.uk/impactofsocialsciences/2018/10/01/replication-is-both-possible-and-desirable-in-the-humanities-just-as-it-is-in-the-sciences/.

Peels, Rik, and Lex Bouter. 2018b. "The Possibility and Desirability of Replication in the Humanities." Palgrave Communications 4 (95). https://science.sciencemag.org/content/349/6251/aac4716.

Pender, Bart, J. Britt Holbrook, and Sarah De Rijke. 2019. "Rinse and Repeat: Understanding the Value of Replication across Different Ways of Knowing." Publications 2019 7 (3): 1-15. https://ideas.repec.org/a/gam/jpubli/v7y2019i3p52-d249307.html.

Schöch, Christof. 2017. "Wiederholende Forschung in Den Digitalen Geisteswissenschaften." In Konferenzabstracts DHd2017: Digitale Nachhaltigkeit, edited by DHd-Verband. https://doi.org/http://doi.org/10.5281/zenodo.277113.

Sinclair, Stéfan, and Geoffrey Rockwell. 2015. "Epistemologica." 2015. https://github.com/sgsinclair/epistemologica.

Sutton, Adam, Thomas Lansdall, and Nello Cristianini. 2018. "Biased Embeddings from Wild Data: Measuring, Understanding and Removing." In Advances in Intelligent Data Analysis XVII. https://doi.org/https://doi.org/10.1007/978-3-030-01768-2_27.

Underwood, Ted. 2019. Distant Horizons: Digital Evidence and Literary Change. Chicago: The University of Chicago Press.

Notes

1. See: https://nips.cc/Conferences/2019/Committees.

2. For a selection of responses, see relevant contributions to Cultural Analytics: https:// culturalanalytics.org/articles?tag=replication

3. For an earlier iteration of the typology, see Schöch (2017). For typologies in other fields, see Gómez, Juristo, and Vegas (2010) and Hüffmeier, Mazei, and Schulte (2015).

4. A practice-oriented guide is Deckker and Lackie (2016).

5. See: https://www.nwo.nl/en/research-and- results/programmes/replication+studies. Call for applications: https://www.nwo.nl/en/funding/our-funding- instruments/sgw/replication-studies/replication- studies. html.

6. The discussion was continued in (Peels and Bouter 2018b), Peels and Bouter (2018a) and in Pender , Holbrook, and De Rijke (2019).

7. See: https://www.nwo.nl/en/news-and-events/news/2019/03/third-round-in-pilot-replication-studies-now-includes-the-humanities.html.

8. See: https://www.nwo.nl/en/research-and- results/programmes/magw/replication-studies/ awards-2019.html.

Constructing Spatial Narratives: Considerations and Practices Across Communities

Christy Lynn Hyman
christy@huskers.unl.edu
University of Nebraska Lincoln, United States of America

Alli Crandell
allicrandell@gmail.com
Coastal Carolina University, United States of America

Sue Bergeron
sbergero@coastal.edu
Coastal Carolina University, United States of America

Jesse Rouse
jesse.rouse@uncp.edu
University of North Carolina-Pembroke, United States of America

Shane Lynch
shane_lynch@ku.edu
University of Kansas, United States of America

Jamila Moore-Pewu

jmoorepewu@fullerton.edu
California State University-Fullerton, United States of America

Bryan Carter
University of Arizona, United States of America

Hilary Green
University of Alabama, United States of America

Keywords: digital mapping, public history, critical GIS, social justice, recovery
Topics: Comparative (2 or more geographical areas), English, 5th-14th Century, 19th Century, Contemporary, public humanities collaborations and methods, spatial & spatio-temporal analysis, modeling and visualization, Geography and geo-humanities, History

This panel considers how critical, innovative approaches with GIS draw our attention to new pathways of digital mapping. How can the practice of digital mapping within a critical lens produce new cartographies for spaces of possibility? How can spatial narratives restore the tarnished lineages of cultural geographies obscured from history?

We want to engage with these questions in our panel. Our hope is to highlight the ethical responsibilities of critically engaged mapping projects. The assorted projects that we propose to present on in this panel are connected through a theoretical grounding that harnesses the power of imagining cultural recovery of landscape as a means of redress from historical obscurity. Ultimately our panel seeks to challenge the matrix of colonial epistemic power underlying traditional foundations of how we construct spatial constructions of communities.

Global Renderings in the Queer Digital Humanities

Tully Barnett
tully.barnett@flinders.edu.au
Flinders University, Australia

Megan Cytron
mcytron@ucm.es
Universidad Complutense de Madrid

Rahul Gairola
Rahul.Gairola@murdoch.edu.au
Murdoch University/ Asia Research Centre

Tyne Daile Sumner
tdsumner@unimelb.edu.au
University of Melbourne

Keywords: queer, intersectional, textual, margins, surveillance
Topics: Europe, English, North America, Australia/ Oceania, 20th Century, Contemporary, digitization (2D & 3D), text mining and analysis, Gender and sexuality studies, Literary studies

How might DH textual scholars respond to call to queer DH practices, methodologies and projects? This panel brings together four papers investigating the queer margins of digital textual or methodological culture. Barnett considers the queer politics of metadata in digitisation projects identifying the way item-level markings become work-level interpretive interventions. Cytron deploys close and distant reading to position Eduardo Mendicutti's 1982 radical act of queer world-building in a global context through analytical stylometry and TEI. Sumner considers the queer implications of distant reading techniques on the FBI files on Langston Hughes and James Baldwin, which conflate homosexuality, race, and poetics by identifying all three as "perversity." Gairola deploys queer of color critique as a lens to read the intersectional nexus of queer and postcolonial theory through analyzing machine learning's supposed ability to detect "gayface" with facial recognition software. Together, these papers identify and press upon diverse margins in global queer digital inscriptions.

Digital Humanities Methods for Analyzing Holocaust and Genocide Testimonies

Todd Presner
presner@ucla.edu
UCLA, United States of America

Anna Bonazzi

annabonazzi@g.ucla.edu
UCLA, United States of America

Lizhou Fan
lizhou@ucla.edu
UCLA, United States of America

Gabor Toth
gabor.toth@maximilianeum.de
University of Southern California, United States of America

Rachel Deblinger
rdeblinger@library.ucla.edu
UCLA, United States of America

David Shepard
dave@humnet.ucla.edu
UCLA, United States of America

Keywords: Holocaust and genocide studies, testimony, text analysis, visualization
Topics: Comparative (2 or more geographical areas), Global, Europe, English, 20th Century, data modeling, text mining and analysis, History, Literary studies

The purpose of this panel is to present a set of digital methods for analyzing Holocaust and genocide testimonies at scale. The corpus of testimonies (approx. 55,000) and metadata come primarily from the USC Shoah Foundation and include survivors of the Armenian Genocide (1914-23), the Nanjing Massacre (1937/38), the Holocaust (1939-45), and the Rwandan Genocide (1994). Panel members also work with early Holocaust testimonies recorded in Displaced Persons Camps by David Boder (1946) as well as testimonies of from Yale's Fortunoff Archive. One of the central research questions concerns the genre of "testimony" itself and how computational analysis can help us track changes in narrative structure, form, and content, particularly in dialogical interviews. We are also interested in how testimonies can be textually mined to fill in "gaps" and "missing voices" through linguistic analysis, including code switching, speech patterns, changes in voice and emotional expressivity, and so forth.

Compartir lo que nos une. Digitizing and Curating Colonial Records from the Caribbean and Central and South America for Public Outreach

Hadassah St. Hubert
hsthuber@fiu.edu
CLIR Postdoctoral Fellow-Digital Library of the Caribbean at Florida International University, USA

Antonio Rojas Castro
antonio.rojas-castro@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften (BBWA), Germany

Tobias Kraft
Berlin-Brandenburgische Akademie der Wissenschaften (BBWA), Germany

Kathrin Kraller
amalia@heritedge.foundation
Berlin-Brandenburgische Akademie der Wissenschaften (BBWA), Germany

María José Afanador-Llach
mj.afanador28@uniandes.edu.co
Universidad de los Andes, Colombia

Amalia S. Levi
The HeritEdge Connection, Barbados

Keywords: colonialism, Caribbean, slavery, digitization, curation
Topics: South America, Comparative (2 or more geographical areas), Europe, English, Spanish, 18th Century, 19th Century, digital libraries creation, management, and analysis, History, Library & information science

Compartir lo que nos une. Digitizing and Curating Colonial Records from the Caribbean and Central and South America for Public Outreach

Overview

When digitizing and curating the colonial past, digital objects are not simply surrogates, but also ways of seeing and interpreting the world: slavery, colonization,

and transculturation are embedded in any document, and if interrogated, their digital representations can become tools for understanding the impact of imperialism and colonialism in the present.

Consequently, digitization and curation of nineteenth-century Caribbean and Central American collections give rise to new questions and challenges in the field of DH: How can we preserve endangered documents while keeping in mind their present-day needs & concerns? How can we digitize and give access to documents dispersed across international institutions to create meaningful collections around slavery? How can we curate data to imagine and visualize space beyond--often artificially imposed--geographical and temporal constraints? Or how can create digital records about oppressed people that transcend their oppressors?

This panel brings together a diverse group of researchers from the Caribbean, South America, Central America, and Europe in order to place digitization and curatorial practices within the analytic framework of cultural encounter described as "post-colonial computing" (Dourish, 2010: 91) and of the post-custodial institutional partnership model (Kelleher, 2017; Alpert-Abrams, Bliss, & Carbajal, 2019). While these encounters and partnerships enable new and exciting collaborations, their processes, epistemological conditions, and products should be critically examined in order not to replicate neo-colonial attitudes often ingrained in digital technologies, and instead support local communities and promote public outreach.

The panel addresses one of the main themes of the conference--"the historical and continued impacts of colonialism, postcolonialism, and hegemony"--and talks will be both in English (Hubert and Levi) and Spanish (Rojas Castro, Kraft and Kraller, and Afanador-Llach); presenters will provide translations to facilitate dialogue between participants and audience.

Protecting Haitian Patrimony Initiatives at the Digital Library of the Caribbean (dLOC)

Hadassah St. Hubert
CLIR Postdoctoral Fellow
Digital Library of the Caribbean at Florida International University hsthuber@fiu.edu

How can we protect and preserve endangered archives in Haiti? In 2010, Digital Library of the Caribbean (dLOC) launched the Protecting Haitian Patrimony Initiative, which builds on strong, existing long term partnerships, with an emphasis on accountability and transparency. This talk will focus on the numerous digitization projects, digital exhibits created since the launch of the initiative, and how dLOC engages the public.

dLOC built Haiti: An Island Luminous, a tri-lingual website to help readers learn about Haiti's history. "Haiti: An Island Luminous" combines rare books, manuscripts, and photos scanned by archives and libraries in Haiti and the United States with commentary by over one hundred (100) authors from universities around the world. "Haiti: An Island Luminous" contextualizes hundreds of historical books, documents and photos digitally preserved by dLOC's partners, including the National Archives of Haiti, the National Library of Haiti, Haitian Library of the Fathers of the Holy Spirit (Pères du St-Esprit), Haitian Library of the Brothers of Christian Instruction (Frères de l'Instruction Chrètienne), University of Florida, Brown University, and University of Central Florida.

dLOC currently hosts over 40,000 titles with more than four million pages of content, much of which is accessible content related to Haiti. Since the launch of "Haiti: An Island Luminous", exhibit stations were placed in the Little Haiti Cultural Center, Nova Southeastern University's Museum of Art, Sant La Haitian Neighborhood Center and more recently at North Miami Public Library in an effort to reach an even larger audience.

dLOC partners in Haiti have been able to contribute more digital content due to the efforts of many scholars that have collaborated to apply for Endangered Archive Programme grants, as well as support from the World Bank, the U.S. Embassy in Haiti, and the Haitian Studies Association. All of these efforts have made dLOC the largest open access repository of Caribbean content and a significant resource for finding materials from and about the Caribbean for use in teaching, research, cultural and community life.

Reconstruyendo la huella de Humboldt en Cuba. Retos y oportunidades de la digitalización del patrimonio documental cubano-alemán del siglo XIX

Antonio Rojas Castro, Tobias Kraft y Kathrin Kraller
Berlin-Brandenburgische Akademie der Wissenschaften
antonio.rojas-castro@bbaw.de
kraft@bbaw.de
kathrin.kraller@bbaw.de
Grisel Terrón, Eritk Guerra y Alaina Solernou
Oficina del Historiador de la Ciudad de La Habana
grisel@patrimonio.ohc.cu
eritk@dic.ohc.cu
alaina@patrimonio.ohc.cu

Con esta comunicación pretendemos debatir sobre los principales retos y oportunidades que supone la cooperación, en el marco del Proyecto Humboldt Digital (ProHD), entre la Academia de las Ciencias y las Humanidades de Berlín (BBAW) y la Casa Humboldt de la Oficina del Historiador de la Ciudad de La Habana

(OHCH). En concreto, nos gustaría responder a la siguiente pregunta: ¿cómo podemos cooperar para el beneficio mutuo (Sennett, 2012)? A continuación, analizaremos el estado actual de la digitalización y las Humanidades Digitales en Cuba. Por último, expondremos los primeros avances orientados hacia la construcción de un repositorio digital y la edición de documentos.

El futuro repositorio digital de ProHD tiene por objetivo preservar y dar acceso en línea a varias colecciones de documentos de naturaleza transnacional y multilingüe (español, francés y alemán), que permitan reconstruir no solo la "huella" del científico prusiano, sino también el pensamiento de muchos intelectuales y políticos locales, como Francisco Arango y Parreño, y las condiciones en que se producía el negocio esclavista, por ejemplo, mediante la digitalización de facturas de compra y venta. De esta manera, al poner la figura de Alexander von Humboldt en su contexto histórico, los usuarios obtendrán una mirada nueva y completa al problema de la esclavitud y al funcionamiento del sistema colonial, que condenaba a Cuba al monocultivo y a importar bienes de primera necesidad procedentes de la Metrópolis.

A fin de obtener digitalizaciones y ediciones académicas digitales fáciles de encontrar, accesibles, interoperables y sostenibles, se han planeado tres medidas principales, que están implementándose desde junio de 2019: en primer lugar, se ha adquirido equipamiento para reforzar la infraestructura tecnológica (ordenadores, escáneres de alta resolución, servidor, impresora, mobiliario, etc.) de la Casa Humboldt en La Habana en donde se llevará a cabo la digitalización de los documentos; en segundo lugar, se ha formado a los miembros del equipo mediante cursos (en línea, sobre todo a partir del confinamiento provocado por la COVID-19) sobre el proceso de digitalización, la creación de metadatos y la gestión de repositorios digitales. Por último, se han definido de manera conjunta unos criterios de selección de documentos, una política de digitalización y un flujo de trabajo compartido.

Inventar el virreinato de la Nueva Granada: curaduría crítica de fuentes primarias y la construcción de conjuntos de datos espaciales, 1739-1810

María José Afanador-Llach
Universidad de los Andes, Colombia
mj.afanador28@uniandes.edu.co

La invención del virreinato del Nuevo Reino de Granada (hoy Venezuela, Colombia, Ecuador y Panamá) en 1739 respondió a un proyecto económico de la monarquía española para proteger el norte de Suramérica de incursiones extranjeras y extraer mayores recursos para la corona. Dicho proyecto de búsqueda de unidad administrativa y geográfica se nutrió de la producción de conocimiento sobre economía política, es decir, sobre cómo crear riqueza en contextos de competencia. En fuentes primarias como mapas, textos sobre economía y descripciones geográficas creados por burócratas, naturalistas y militares en la colonia, se evidencian las dificultades de integrar un territorio montañoso y extenso y la importancia de la imaginación geográfica en este proceso.

El proyecto de integración territorial del virreinato se rompió con la crisis monárquica de 1808 cuando el reino se fragmentó en más de una docena de provincias autónomas. La crisis representó una oportunidad para que pueblos y ciudades avanzaran sus proyectos económicos locales y se imaginaran espacios económicos post-coloniales. La pregunta central que guía este proyecto es: ¿cómo se puede visualizar la relación entre la búsqueda de unidad territorial en el norte de Suramérica y la diversidad de paisajes y economías políticas en tensión en una interfaz de archivo y un mapa digital?

Esta ponencia explora los avances de este proyecto en sus dos primeras fases. La primera fase consiste en la curaduría crítica de fuentes primarias entre 1739 y 1810. Esta curaduría consiste en la selección de documentos, de fragmentos dentro de los documentos y de mapas del periodo, para organizarlos, enriquecer sus metadatos y disponerlos alrededor de un argumento y una narrativa histórica. Algunos de estos documentos se encuentran disponibles en libros editados, otros están digitalizados y un número pequeño se digitalizará. La segunda fase consiste la formulación de un proceso metodológico para extraer datos sobre las dimensiones espaciales y de economía política de este conjunto curado de fuentes primarias. Se trata de generar conjuntos de datos relacionales que permitan visualizar, analizar e interpretar transformaciones históricas-espaciales de la invención y crisis del virreinato de la Nueva Granada y los actores involucrados en este proceso. El proyecto busca que tanto la curaduría de fuentes primarias como los conjuntos de datos sean recursos abiertos en español para la investigación y docencia sobre el periodo. La ponencia discutirá los retos a los que se ha enfrentado el proyecto hasta el momento en términos de la curaduría de documentos y la definición metodológica para la extracción de datos.

Beyond digitization: Engaging the Community to Decolonize the Archival Record for the Enslaved

Amalia S. Levi
Archivist, Chair
The HeritEdge Connection, Barbados
amalia@heritedge.foundation

During 2018 and 2019, two historic newspapers, The Barbados Mercury Gazette (1783-1848) and The Barbadian (1822-1861), housed in the Barbados Department of Archives, were digitized through Endangered Archives Programme grants. As primary sources, they offer a detailed view of every aspect of the dystopian life in a British colony in the Caribbean during the eighteenth and nineteenth centuries. Digitization offers unparalleled access to these colonial newspapers that were previously largely inaccessible due to their fragility.

The presentation goes beyond digitization to discuss what happens after digitization is completed. Considering that the newspapers were part of the colonial information apparatus, how do we read against the grain to locate and reveal marginalized voices hidden in the digitized pages?

If our aim is to "lift" the voices of the enslaved from the pages, and retrospectively create a body of archival records, we need to provide access to the digitized pages as data. Such information however exists "locked" in digitized images that due to the deterioration of the paper and the discoloration of the pages are impossible to OCR. Finding and transcribing these ads is necessarily a manual process, at present.

Among the wealth of information appearing regularly in the newspapers, of particular importance are the "runaway slaves" ads. Archives usually preserve very little descriptive information about the enslaved, because their lived experiences were rarely recorded. When they appear in the archival record, they are enumerated or appraised as commodities. Thus these ads offer a rich trove of information about individual people, including name, age, physical appearance, skin color, clothing, accent, distinguishing features (such as body modifications from their country of origin or bodily harm, the result of violence), friends, relatives, and skills.

Digitization does not solve the gaps, silences, and omissions inherent in the archival record and the colonial epistemologies they contain. Colonial newspapers reflect the voices of the white class of planters, merchants, and colonial authorities. Simply digitizing and putting online risks elevating and amplifying the very colonial worldviews we mean to avoid.

To decolonize the record about the enslaved, digitization is only the starting point for further projects and initiatives to engage the community with its own history. Recent scholarship and public humanities efforts have shown that it is possible to challenge the erasure of colonial archival sources and read between the lines to tease out information that is not readily visible.

Our aim is to create a collection of material by clipping "runaway slaves" ads from the digitized newspapers, transcribing the text, and enriching the human stories in each ad with additional contextual information. More specifically, the first part of the presentation will focus on work to develop the "Barbados Runaway Slaves Digital Collection," a partnership between the Barbados Archives, the local non-profit HeritEdge Connection, and the Early Caribbean Digital Archive (ECDA) at Northeastern University (Boston).

This digital collection aims to foreground the centrality of enslaved voices by 'lifting' ads of individuals who chose to escape slavery from the pages of the newspapers, and turning them into individual, standalone archival records previously unrecorded in the archives. When completed, the "Runaway Slaves Digital Collection" will provide a central location for collecting and presenting these ads, and other opportunities for the public, especially students, both in Barbados, as well as abroad, to interact with the material in creative ways.

By 'reading' these ads in various ways and being able to 'see' them collectively, we can start seeing patterns and coping mechanisms. At the same time, the availability of this information points to what is not there, and invites us to be sensitive to gaps and silences.

The second part of the presentation will discuss public outreach initiatives. Material digitized through generous grants by institutions in the Global North are eventually hosted in digital platforms in those institutions. While praiseworthy, digitization through such grants ends up benefiting scholars in the Global North. Usually local people are unaware of these platforms, and often unable to access them, either due to bandwidth issues or simply to interfaces that might not be intuitive to use. We have tried to remedy this by focusing on public history work we are doing to increase awareness of and engagement with these digitized colonial records.

During the fall of 2019, we conducted a series of workshops aimed to familiarize the public with accessing the newspapers online and to transcribe ads. The workshops also provided a platform for the public to discuss the ads, and the many facets of slavery. People were able to see ancestors who chose to resist and escape bondage in adverse, inhumane conditions. Due to COVID-19, workshops planned for 2020 are being held online. The aim of these workshops is to engage people with the ads in creative ways, through genealogical research, speculative writing, or digital methods. In this way, we invite the public to contextualize the ads through their local knowledge of places mentioned in the ads, or intimate information

about lived experiences. Beyond information that is there, we also hope that workshop participants can imagine what is not there, complete the stories, and give enslaved individuals their place in the archival record.

Additionally, the digitization of these primary sources has great potential for digital projects by students and other researchers that can highlight various aspects of the island's history. The "runaway slaves" ads help people challenge the customary narrative of Barbadian passivity and submissiveness to slavery; reconstruct family and community networks that supported enslaved Barbadians; and they help give voice to ancestors, whom colonial records intentionally left voiceless.

Bibliographic References

Alpert-Abrams, H. Bliss, D. A., and Carbajal I. "Post-Custodialism for the Collective Good: Examining Neoliberalism in US-Latin American Archival Partnerships". In: "Evidences, Implications, and Critical Interrogations of Neoliberalism in Information Studies", eds. Marika Cifor and Jamie A. Lee. Special issue, Journal of Critical Library and Information Studies 2.1 (2019). DOI: https://doi.org/10.24242/jclis.v2i1.87

Dourish, P. "'Computational Thinking' and the Postcolonial in the Teaching from Country Programme". Learning Communities. International Journal of Learning in Social Contexts, 2 (2010): 91-101.

Kelleher, Ch. "Archives without Archives: (Re)Locating and (Re)Defining the Archive Through Post-Custodial Praxis". In: "Critical Archival Studies", eds. Michelle Caswell, Ricardo Punzalan, and T-Kay Sangwand. Special issue, Journal of Critical Library and Information Studies, 1.2 (2017). DOI: https://doi.org/10.24242/jclis.v1i2.29

Sennett, Richard. Juntos. Rituales, Placeres y políticas de la cooperación. Barcelona: Anagrama, 2012.

Risking I.T.: Trans-Feminist Queer Digital Performance, Archives and Information Technologies

Jas Rault
jas.rault@utoronto.ca
University of Toronto

TL Cowan
tl.cowan@utoronto.ca
University of Toronto

Naveen Zehra Minai
naveenminai@gmail.com
University of Toronto

Carina Guzmán
cartogeosapiens@gmail.com
University of Toronto

Henria Aton
henria.aton@mail.utoronto.ca
University of Toronto

Keywords: trans feminist queer, ethics, translocal, anti-colonial, archives
Topics: Asia, Comparative (2 or more geographical areas), English, North America, 20th Century, Contemporary, cultural analytics, digital archiving, Cultural studies, Gender and sexuality studies

This panel brings together research of the Trans Feminist Queer Digital Praxis Workshop. Cowan and Rault will open the panel with a presentation outlining the "heavy processing" ethico-methodologies that drive our DREC and Cabaret Commons projects. Minai will speak to "three digital sites of queer and trans female masculine fashion in the United States to think about non-normative masculinities in digital visual cultures." Guzmán will take up the "queer impulse that led [Mexico City based] Meras efímeras to imagine Machistán has become an effort to digitize and archive materials and collect memories about our party-organizing political endeavor on a digital platform." Aton will discuss "three feminist digital archives related to human rights in Sri Lanka: the Herstories archive, Adayaalam Centre for Policy Research's 'Stories of Resilience' archive, and one hypothetical, unnamed archive that does not yet exist but that would include materials and narratives around enforced disappearances."

Working at the Intersection: Digital Humanities Pedagogy, Social Justice, and Librarianship

Pamella R. Lach
plach@sdsu.edu
San Diego State University

Megan Martinsen
megan.martinsen@georgetown.edu
Georgetown University

Amanda Visconti
visconti@virginia.edu
University of Virginia

Brandon Walsh
walsh@virginia.edu
University of Virginia

Heather Froehlich
hgf5@psu.edu
Penn State University

R.C. Miessler
rmiessle@gettysburg.edu
Gettysburg College

Nabil Kashyap
nkashya1@swarthmore.edu
Swarthmore College

Keywords: pedagogy, social justice, libraries
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, digital activism and advocacy, Library & information science

At many institutions, the library is at the center of Digital Humanities work on campus. Academic libraries are important discipline-spanning partners in the digital humanities, from research to teaching. DH library staff engage in a broad range of pedagogic activities, including teaching for-credit classes, partnering with teaching faculty in their classrooms, and offering workshops, to name a few. These sites of learning offer possibilities for social justice pedagogy across a spectrum of pedagogic praxis. This panel will explore a range of approaches DH librarians and staff across the United States employ for social justice pedagogy, representing a range of institutions from research universities to liberal arts colleges. We will examine ways of infusing critical pedagogy into DH library instruction, including one-shot instruction and embedded librarianship, as well as DH instruction designed specifically for classes related to social justice. We will cover varying approaches to building academic communities based on collaboration, kindness, generosity, and shared values, and we will discuss user-centered DH teaching and learning that is inclusive and responsive to particular student needs. We will also contemplate the perils and possibilities of engaging in DH and social justice work within and beyond libraries. Libraries are an important site to examine the intersections of DH, social justice, and pedagogy, and we hope this panel will offer multiple windows into the on-going work in this area and move forward both the conversation and the practice around the way we engage with DH learners.

Pamella R. Lach, Digital Humanities Librarian, San Diego State University. "Infusing Critical Digital Humanities Pedagogy in the Library Classroom: From One-Shot Instruction to Embedded Librarianship"

Digital Humanities librarians engage in a range of DH-related instruction, from the "one-shot"-a single and standalone interaction with a class-to embedded librarianship-in which a librarian works closely with a class for the duration of a semester. Librarians striving to infuse critical digital humanities approaches into classrooms that are not their own must be prepared to incorporate critical approaches in both small and large ways. This talk explores the opportunities and challenges for one DH librarian attempting to infuse critical digital humanities into their instruction.

Often faced with a limited amount of time with students, I carefully select DH project examples that have a social justice focus, whether a digital history project that serves as a corrective to the archival silencing of people of color, or a project explicitly designed to advance equity, diversity, inclusion (EDI), and social justice through its digital praxis. I also ask students to explore well-intentioned projects that fall short of these goals-for example, projects that re-objectify and dehumanize individuals through problematic visual representations. In examining a range of projects dedicated to people of color-their

histories and stories-my students and I engage in critical sensemaking that interrogates the motivations of the projects' creators, the assumptions and biases baked into the projects, and the unintended messages being delivered. While I cannot always provide solutions to these problems, the questioning that we engage in together leads to a critical praxis that I hope they will apply to other classes and scholarship (digital or analog).

More often than not, I work closely with classes over an entire semester in a more embedded style, which affords me opportunities for even greater social justice interventions.

When engaging in ongoing work with a class-whether project- or tools-based-I bring in that same critical framework. But my obligation towards critical DH extends beyond the examples I use to teach or the conversations we have in class. Working at a lesser-resourced, teaching-intensive institution where not all students own a laptop or are comfortable with technology, I often select free, lower-tech solutions with minimal access barriers. But in selecting such tools, many of which rely on Google integration, I am also aware of the risks to the students, particularly for undocumented students who might be vulnerable using tools that do not guarantee data privacy or expose them to surveillance risks. While opting for expensive tools may not necessarily be the solution, I try to make these choices transparent, call attention to the risks, and offer meaningful alternatives and opt-out solutions so that students can make informed decisions about how they want to engage in DH in their classes.

Even when we are not instructors of record, DH librarians can make important contributions to the classroom by modeling critical inquiry and helping students evaluate digital scholarship from an EDI- and social justice-informed lens. Whether they are reviewing existing digital projects or creating their own, we can teach students to make careful-if constrained-choices, and to be open and honest about those choices.

Megan Martinsen, Digital Scholarship Librarian, Georgetown University. "Digital Humanities Methods for Social Justice Pedagogy"

The growing discussion about infusing digital humanities pedagogy with social justice concepts is incredibly important work and exciting to see. Less discussed, but equally exciting in my view, is the reverse. Embedding digital humanities tools and methods into social justice pedagogy is a useful and important endeavor. It can open up avenues for social justice students to share their important work online as public scholars and as students gain comfort with digital research methods, they can potentially ask new questions about the social justice topics they study. My talk will present a case study of a social justice course at my institution where I collaborated with the teaching faculty member to create a final project that used digital humanities methods and where I served as an embedded librarian for the semester. The course, Justice and Peace Studies 299, is the research methods course required for the Justice and Peace Studies major and minor. The bulk of the course is spent exploring research methods used in social justice studies. My collaboration with the teaching faculty member added a section on digital humanities methods that can be applied to social justice studies.

The assignment we created had students work in groups focused on a social justice issue of their choosing. Each group was required to write a paper that employed basic text mining, create an infographic, create a data visualization, and create a map. My talk will delve into how the assignment came to be, student reactions to the assignment, and common issues the student groups faced. And I'll discuss lessons learned from the project and how we might change and shift the project in upcoming semesters.While not every institution has a social justice degree path, they are growing in popularity and many institutions now offer at least some coursework in social justice studies. My hope is that this talk will help digital humanities practitioners see a role for themselves as collaborators with these emergent programs.

Brandon Walsh, Scholars' Lab Head of Student Programs, University of Virginia Library; Amanda Visconti, Scholars' Lab Managing Director, University of Virginia Library. "Building Community and Generosity in the Context of Graduate Education"

The academy trains students in complex intellectual work and tends to reward the performance of one's own intelligence, both in coursework and in conferences. But in collaborative digital projects, this sort of focus on the individual is detrimental to the group dynamic, which must take shape around more than just the individual. In our year-long introduction to digital humanities by way of project-based pedagogy, we consistently noted this tension in our fellowship cohorts. Each year, we struggled to promote healthy collaboration centered on shared buy-in and generosity; our students instead remained focused on what a group project would mean for them and their careers. While these are important needs and understandable desires, we hoped to use their year with us as an opportunity for them to grow beyond this framing. We eventually realized that our problem was expecting the students to act in a way different than habitually expected and taught in other parts of the academy. We realized this was a matter of training: our students were trained in seminar-type environments that value the performance

of individual thought, and never were exposed to or rewarded for a different approach. This talk discusses workshops and exercises for our graduate student fellows that encourage reorienting their collaborative practices away from a focus on the self, and towards creating more generous and kind community spaces. We will focus on a workshop on community and collaboration we now teach for our students, where we discuss what might be gained by de-centering yourself, building up rather than deconstructing, leading from trust and kindness, and more. This session is part of a larger exercise in producing group statements of values-charters-that require students to examine their own personal goals, experiences, and backgrounds in relationship to the whole. This process tries to help students recognize that these conversations deserve a space in academic discourse, that academic practice can be different than they've experienced elsewhere, and that this conversation can be a space in which to begin to shape a different kind of scholarly work.

Heather Froehlich, Literary Informatics Librarian, Penn State University. "Digital Pedagogy at Scale: Reaching Different Student Populations"

How we adapt classroom pedagogy based on the needs of a given student population is a common discussion in instruction and pedagogy circles. However, this has been less of a discussion in Digital Humanities instruction: the assumption seems to be primarily focused on a specific population signed up for a class with a DH element. But what happens when your institutional population covers a range of students and socioeconomic backgrounds across a wide range of the state's demographics? In this talk I will discuss some practical considerations about teaching digital methods across a multi-campus state university system using two examples of digital humanities instruction related to early modern digital projects. One project at a well-resourced flagship campus involving a graduate seminar in scholarly editing where students are learning about digital scholarly editing with TEI; the associated undergraduate class is learning to flex their muscles in a lower-stakes editing task. The other project, at one of our 4-year undergraduate, more residential colleges is an exercise in undergraduate student research related to 17th century recipe books; students transcribe and are developing research projects around this manuscript.

In each situation, the pedagogical contexts are wildly different: at one campus, our students are widely understood to be from more upper middle-class families and often have a parent or family member who went to college; at another campus, our students are primarily local and a significant population are 1st-generation and/or coming in via local, urban community colleges. Our graduate students represent a third vector, given that they all have undergraduate degrees and many of them are in a combined MA-PhD program. Developing models of instruction for these classes require a level of flexibility for the librarian and faculty partners: what works for one group does not always work for all others.

R.C. Miessler, Systems Librarian, Gettysburg College. "Values-Based Communities of Practice in the Digital Humanities"

The Digital Scholarship Summer Fellowship is a library-led, high-impact learning experience for undergraduates at a liberal arts college that provides students an introduction to Digital Humanities tools and methods. Student fellows receive a stipend for the summer to participate in the program, in which they choose their own research question and create a digital project. While the students' research topics may or may not share themes, tools, or methods during the course of the summer program, all fellows participate in a common curriculum, read and discuss a variety of readings from a diverse range of DH practitioners, and publicly present their projects at the end of the summer. The librarian facilitators of the program draw from their own areas of expertise or interest to mentor students, deliver workshops, and guide discussions. Together, both students and librarians broaden their own understanding of the nature and workings of DH in a space that encourages experimentation and failure.

The core experience of the summer program is the emphasis on the development of a community of practice amongst the student cohort and librarian facilitators; rather than focusing on a single research question or method, the community is centered on shared values. To support a wide range of research questions, the librarians who facilitate the fellowship approach Digital Humanities pedagogy from a generalist and discipline-agnostic perspective when teaching tools and methods. With each student fellow undertaking an independent research project, workshops focus on critical evaluation of the affordances and limitations of the Digital Humanities more broadly, as well as introducing students to project management, user experience, and public speaking. Through the framework of the curriculum, students have agency to choose which tools and methods are best suited to their research while acknowledging the limitations of what they can reasonably accomplish in the course of the summer program.

This talk will focus on the evolution of the development of local, values-based DH communities of practice that transcend typical administrator/ undergraduate hierarchies and how those values are transmitted and reinforced via experimentation, mentorship, and pedagogy. Past student projects

created during the summer programs that have focused on the local community and issues of social justice will also be highlighted to show how the values have carried on in student research activity.

Nabil Kashyap, Digital Scholarship Librarian, Swarthmore College. "Institutional Tactics: Locating Social Justice in Liberal Arts DH"

Drawing on a breadth of library activities at a liberal arts college, this talk explores how our campus has engaged highly elusive, tangled issues around digital humanities pedagogy and social justice. The questions that animate our work: Where can we find and make space in our professional practices to address injustice? How can we make that work legible and articulate value when it takes place diffusely across a range of decisions, sites, and stakeholders? Given our context, how can we avoid being subsumed by trends in higher education and resist institutional virtue signaling?

We examine three loci of recent activity: faculty pedagogy, student programs, and ad hoc collaboration. We regularly design and lead workshops aimed at instructors around digital assignments and implementing DH tools in the classroom. The resulting skills intensives have taken place both at liberal arts colleges and larger campuses. This process affords us opportunities to explicitly engage the perils and possibilities of digital assignments and platforms--from the violence of metadata to military/colonial histories of particular tools to minimal web development for decentralized, public humanities.

We run several co-curricular student programs including a digital scholarship student initiative that brings humanities and computer science majors together around theory and praxis of digital environments. We read widely in STS, Post-Colonial DH, and media theory, while working hands-on with the web stack, attentive to the ecological, political, and social dimensions of everyday digital infrastructure.

We also have regular opportunities to collaborate with other units on campus. How might we, for example, partner with initiatives oriented around innovation and entrepreneurship? One way was to co-sponsored a hackathon, which allowed us to shape the structure and goals of event. Through that process we were able to eschew commercial webapps in favor of partnering with local activist organizations representing pressing community needs.

Rather than showcasing disparate library initiatives, these activities taken in aggregate point to the library as a site for contingent tactics deployed alongside institutional strategies. We engage in flexible if fallible approaches to our professional and personal commitments to social justice as enacted through and imperiled by digital environments.

Twitter is an Indigenous Territory

Ashley Caranto Morford
ashley.morford@mail.utoronto.ca
University of Toronto, Canada

Jordan McVittie
jordan.mcvittie@mail.utoronto.ca
University of Toronto, Canada

Katelyn Ward
katelyn.ward@mail.utoronto.ca
University of Toronto, Canada

Nicole Santos Dunn
nicole.dunn@mail.utoronto.ca
University of Toronto, Canada

Samantha McCormick
samantha.mccormick@mail.utoronto.ca
University of Toronto, Canada

Sewsen Goitom-Igbu
sega_sew@yorku.ca
University of Toronto, Canada

Jeffrey Ansloos
University of Toronto, Canada

Keywords: Indigenous studies, hashtag ethnography, language, artistic revitalization, life promotion
Topics: English, North America, Contemporary, digital activism and advocacy, social media analysis and methods, First nations and indigenous studies

#DecolonizingDigital is a project that runs out of the Indigenous Life Promotion & Social Action Network Labs at the Ontario Institute for Studies in Education. The project investigates how Indigenous people are using Twitter for cultural revitalization, well-being, and community connection. This panel offers three papers with findings for stage one of the project. The first paper discusses how Twitter-based Indigenous language revitalization movements enable Indigenous language learners to live their languages in culturally grounded and place-based ways, and present

Indigenous languages not as dying but, rather, as vibrant, active, and living. The second paper discusses Twitter as an Indigenous community shaped by beaders who are connecting and sustaining culture, love, identity, healing, and Indigenous trade systems through beading in Twitter's ecosystem. The third paper discusses how We Matter, an Indigenous youth-led organization, is challenging colonial theorizations of mental health and re-articulating Indigenous health through the lived experiences and knowledges of Indigenous youth.

Cultural Analytics and the Book Review: Models, Methods, and Corpora

Matthew J. Lavin
lavin@pitt.edu
University of Pittsburgh, United States of America

Kent Chang
Carnegie-Mellon University

Yuerong Hu
yuerong2@illinois.edu
University of Illinois

Wenyi Shang
wenyis3@illinois.edu
University of Illinois

Aniruddha Sharma
University of Illinois

Shubhangi Singhal
University of Illinois

Ted Underwood
tunder@illinois.edu
University of Illinois

Jessica Witte
University of Illinois

Peizhan Wu
University of Illinois

Dan Sinykin
daniel.sinykin@emory.edu
Emory University

Melanie Walsh
melanie.walsh@cornell.edu
Cornell University

Maria Antoniak
maa343@cornell.edu
Cornell University

Keywords: book reviews, reception studies, periodicals, distant reading, canonicity
Topics: English, North America, 19th Century, 20th Century, Contemporary, cultural analytics, data publishing projects, systems, and methods, Book and print history, Literary studies

In 2013, Peter Boot highlighted the merits of a corpus of online book responses. "If reading responses are important for the study of literature and its effects," he argued, "it follows that we need to understand them better."[1] A corpus of book responses would also help us see "relationships between the responses and the works that they respond to, in terms of topics and narrative"; for example "what characters and plot developments do reviewers respond to?"[2] Boot called for a sufficiently large, representative corpus that included many book genres, and relevant contextual metadata meeting the standards of open data access and usability. Six years later, scholarship by James F. English, Allison Hegel, Andrew Piper & Richard Jean So, Dan Sinykin, and Jordan Sellers & Ted Underwood (among others) has shown the range of insights that can be drawn from a corpus of book reviews and/or sets of book review corpora.[3] Simultaneously, prominent voices in digital humanities have called on the scholarly community "to consider the nature of ontological gaps and epistemological biases," including the "infrastructures of knowledge-making" on which large scale, computational studies often depend.[4] This panel will present four papers at the forefront of cultural analytics methods and analysis of book reviews.

In "Modeling Bibliographical Information in Historic Book Reviews: Large Scale Applications with Proquest's American Periodicals Series," Matthew J. Lavin evaluates computational methods to classify items as likely book reviews, differentiate single-work reviews from multi-work reviews, and extract from reviews information about the book and author being reviewed.

"Book Reviews and the Consolidation of Genre" (Kent Chang, Yuerong Hu, Wenyi Shang, Aniruddha Sharma, Shubhangi Singhal, Ted Underwood, Jessica Witte, Peizhen Wu) addresses a common doubt about the significance of text analysis by showing that measurements of similarity between literary texts correlate with similarities between their reviews. It then uses this method to trace the consolidation of genre.

In "Reconstructing Consecration in US Literary History, 1965-2000," Dan Sinykin uses social network analysis methods on book reviews to evaluate novels' centrality in a network. It analyzes the publishers most or least likely to offer an author literary success, the impact of gender and race on reception, and the changing literary value of genre.

In "The Crowdsourced 'Classics' and the Revealing Limits of Goodreads Data," Melanie Walsh and Maria Antoniak analyze more than 100,000 Goodreads reviews of the most popular "classics" in order to reveal the widespread use of "classic" as a colloquial literary critical term that is related to but distinct from the "canon." Through the lens of these reviews, the authors investigate how online readers and institutions both old and new collaboratively construct the highly lucrative category of the "classics."

Collectively, these four papers offer new perspectives on corpus development, modeling the book review as an object of study, and analyzing large sets of book reviews using natural language processing, machine learning, and social network analysis methods.

Notes

1. Boot, Peter. "The Desirability of a Corpus of Online Book Responses." Proceedings of the Workshop on Computational Linguistics for Literature. ACL, 2013, 32.

2. Ibid.

3. See, for example, James F. English, et al. "Mining Goodreads: Literary Reception Studies at Scale," https://pricelab.sas.upenn.edu/projects/goodreads-project; Allison Hegel, "Social Reading in the Digital Age." Ph.D., University of California, Los Angeles, 2018. https://search.proquest.com/pqdtglobal/docview/2061506661/E7D5C6D89E2843D0PQ/1; Andrew Piper and Richard Jean So, "Women Write About Family, Men Write About War," The New Republic, April 8, 2016. https://newrepublic.com/article/132531/women-write-family-men-write-war; Dan N. Sinykin "The Conglomerate Era: Publishing, Authorship, and Literary Form, 1965-2007." Contemporary Literature 58, no. 4 (2017): 462-91; Ted Underwood, and Jordan Sellers. "The Longue Durée of Literary Prestige." Modern Language Quarterly 77, no. 3 (September 1, 2016): 321-44. https://doi.org/10.1215/00267929-3570634.

4. Katherine Bode, "Why You Can't Model Away Bias," Preprint, forthcoming from Modern Language Quarterly, 2. https://katherinebode.files.wordpress.com/2019/08/mlq2019_preprintbode_why.pdf; Bonnie Mak, "Archaeology of a Digitization." Journal of the Association for Information Science and Technology 65.8 (2014): 1519, qtd. in Bode 2.

Applied DH: Lessons from Using the Methods and Technology of Digital Humanities to Build Tools and Services for Researchers, Librarians and Curators

Peter Leonard
peter.leonard@yale.edu
Yale University Digital Humanities Lab

Abigail Potter
abpo@loc.gov
Library of Congress Labs

Lotte Wilms
Lotte.Wilms@KB.nl
KB Lab National Library of the Netherlands

Alex Humphreys
alex.humphreys@ithaka.org
JSTOR, United States of America

Keywords: Applied DH, DH labs
Topics: Global, Europe, English, North America, 19th Century, 20th Century, Contemporary, image processing and analysis, natural language processing, Library & information science

The distinction in the sciences between developing new technologies and methods and the applied sciences which seek practical applications of scientific advances is a useful model for the digital humanities. A set of libraries and laboratories are leveraging DH

methods such as natural language processing and machine vision towards practical applications that are immediately useful for researchers, library patrons and other end-users. This panel will feature a diverse set of libraries and laboratories describing case studies in which they have used DH-inspired and DH-derived methods and technologies to build tools that are changing approaches to service development. Case studies will include the creation of a multi-corpus text mining service, computer vision-based visualization tools, a data access service and service model, and a public volunteer program for transcribing documents. These will illuminate DH's' impact outside of the academy, while offering lessons for practitioners that may help to expand this impact.

Sustainability in the Making and Knowing Project: From Scholarly Edition to Community Crossroads

Pamela H. Smith
ps2270@columbia.edu
Columbia University, United States of America

Nick Laiacona
nick@performantsoftware.com
Performant Software, United States of America

Terry Catapano
catapanoth@gmail.com
University of California, Berkeley, United States of America

Naomi Rosenkranz
njr2128@columbia.edu
Columbia University, United States of America

Tianna Helena Uchacz
tianna.uchacz@gmail.com
Columbia University, United States of America

Clement Godbarge
clement.godbarge@fulbrightmail.org
Columbia University, United States of America

Tillmann Taape
tillmann.taape@cantab.net
Columbia University, United States of America

Keywords: Scholarly Edition, Critical Making, Pedagogy, Art, Science
Topics: Global, Europe, English, 15th-17th Century, Contemporary, scholarly editing and editions development, analysis, and methods, sustainable procedures, systems, and methods, Art history, History of science

Abstract

Discoverability and sustainability bedevil DH projects. They suffer from problems of sustainability not just through obsolescence of technology, but by lack of discoverability and use-sustainability is not just a technological problem but also a social one. The Making and Knowing Project's development of a Community Crossroads confronts these issues by fostering a collaborative culture in humanities research through digital tools to share not only data and finished products, but also insights into the work behind the scenes, and the integration of research and pedagogy. We question what it means for a scholarly digital product to be "final," and we broaden the concept of digital sustainability to include not just the technical sense of static data storage, but also a social environment of continued user involvement and dynamic project development. By addressing discoverability, sustainability, and use, the panel responds to the DH2020 themes of the public digital humanities and open data.

The Making and Knowing Project (M&K)

Since 2014, M&K has studied the intersections of craft making and scientific knowing in the early modern period through the lens of BnF Ms. Fr. 640-a little-known compilation in Middle French by an anonymous "author-practitioner" of roughly 1,000 "recipes" detailing instructions for artistic and technological objects. The manuscript grants insights into the process by which artistic creation yielded insights into the workings of nature. It is a rich and unique example of a "how-to" text-a body of literature whose significance is widely acknowledged, but under-investigated. Over the last 6 years, through crowd- and grad-sourcing in workshops, conferences, and courses, M&K has transcribed, translated, encoded, researched, and annotated the 340-page manuscript. In February 2020, it released a digital critical edition of the entire manuscript: Making and Knowing Project, Pamela H. Smith, Naomi

Rosenkranz, Tianna Helena Uchacz, Tillmann Taape, Clément Godbarge, Sophie Pitman, Jenny Boulboullé, Joel Klein, Donna Bilak, Marc Smith, and Terry Catapano, eds., Secrets of Craft and Nature in Renaissance France. A Digital Critical Edition and English Translation of BnF Ms. Fr. 640 (New York: Making and Knowing Project, 2020), https://edition640.makingandknowing.org. M&K will now transform its static Edition into the core of a dynamic Community Crossroads.

Product vs. Process

Despite being defined as modes of scholarship, the digital humanities are often solely represented by products rather than the methodologies employed in their creation; once the collaborative work ceases, the experiential insights are effaced in the finished project and lost to makers of future projects. While the digital humanities offer new modes of dissemination and sustainability of scholarly products, they also provide opportunities for evaluating, refining, and sharing the research process. "Product" also characterizes the scholarly format of a critical edition, which is intended as a definitive version of a manuscript or manuscript corpus. This focus on definitive product works at cross-purposes to one of the greatest potentials of the digital humanities: user-directed, open-ended interactivity that allows scholarship to be built upon and broken down, with the data, platform, and methods reused and reassembled by reader-users in new and inventive ways.

A recognition of the tension between the Project's aim of the definitive critical edition product of Ms. Fr. 640 and the digital humanities process we employed came to a head as students sought to reconstruct the technical "recipes" of the manuscript in a laboratory course, and we came to see that a recipe text is, by its nature, not a definitive text-its processes can never be fully realized in writing, and they are often much harder to comprehend in words than by doing (or by watching, as YouTube has demonstrated), thus extension into the physical and digital realm was necessary. Moreover, Ms. Fr. 640 is a compiled text, and it can be broken down and rearranged in any number of ways, thus forming an ideal text for the reassemblage and reuse that digital format makes possible. The Project aims to foster new uses and new interpretations of this assemblage of recipes through making its components available in a Community Crossroads.

The Community Crossroads

A series of workshops and courses focusing on pedagogy, research, and outreach will create a community of reader-users of the Edition who take away the Project's teaching resources, the digital infrastructure for its Edition, and its methods and process, to reuse and reassemble in their own diverse settings-whether digital humanities labs, classrooms, museums, or public programs. Fostering the community, and the new and anticipated uses and discoveries it will engender, will extend and sustain-in the fullest sense of the word-the Edition. The Edition's methods and process, hidden by its format as a critical edition, will be made visible and disseminated. The Community Crossroads comprises three main components:

- The M&K "Research and Teaching Companion" will refine and extend the methodologies developed through the creation of its Edition of Fr. 640 and articulate them in a formal implementation guide. The web-based Companion will offer a flexible and adaptable model with resources that scholars, instructors, and students can employ in their own research endeavors or in the classroom. It forms a scalable template for applying and transferring the Project's key methods: interdisciplinarity and collaboration; pedagogy-driven research and research-driven pedagogy; material literacy through historical reconstructions of past techniques; and the harnessing and utilization of the digital humanities to serve research questions.
- The "myEdition" community version will make available an open-source, customizable version of the Edition infrastructure, releasing a community software platform in which others can create their own digital critical editions, using the Project's framework to work with their own data.
- The "Sandbox Space" will be an interactive online environment for others to 1) explore and experiment with the Project's data using a variety of digital tools, and 2) add and create their own tools, case studies, and resources resulting from the use of the Template components.

The Community Crossroads thus aims to overcome and rethink the most serious challenges in digital humanities projects- discoverability, sustainability, and the effacing of the creation process.

Panel Speakers

NAME: Pamela H. Smith
AFFILIATION: Columbia University
EMAIL: ps2270@columbia.edu

TITLE: Insights from the Making of a Digital Edition

ABSTRACT: The Making and Knowing Project's five-year initiative to transcribe, translate, and encode the text of BnF Ms. Fr. 640, a sixteenth-century how- to text with almost 1000 entries for a diverse range of art and technical processes, has yielded evidence about the relationship between craft and science in early modern Europe. It has also provided insights about methodology in collaborative work across humanistic studies, the value of integrating hands-on materials-based and digital learning into curricula, and the viability

of a model of pedagogy-driven lab-based research outside of the hard sciences.

NAME: Nick Laiacona
AFFILIATION: Performant Software
EMAIL: nick@performantsoftware.com

TITLE: A Case Study in Developing Responsive, Functional Design within Minimal Computing Constraints

ABSTRACT: To ensure its archival preservation by Columbia University Libraries, the Making and Knowing Project required that its Digital Critical Edition be developed as a static site while meeting certain core expectations for functionality. This presented several design challenges for Performant Software, including the need for a responsive layout to reconstruct the text and marginalia of the original folio, the adaptation of Lunr's basic functionality to generate satisfying searches of middle French transcriptions and contemporary English translation, and the leveraging of the Project's custom semantic markup toward a meaningful navigation tool.

NAME: Terry Catapano & Naomi Rosenkranz
AFFILIATION: UC Berkeley; Columbia University
EMAIL: catapanoth@gmail.com; njr2128@columbia.edu

TITLE: Social and Technical Strategies for Sustainability

ABSTRACT: The Making and Knowing Project's Edition, conceived within Columbia's ethic of minimal computing, has approached sustainability from two directions: technical and social. While the former led to the development of a reduced-functionality, "small-stack" website to minimize dependencies and avoid rapid obsolescence, the latter has mitigated some of the technical limitations by fostering a "Community Crossroads" where public users can interact with the Project's core data along designated pathways through alternative interfaces (e.g., Jupyter notebooks) or independently through full data downloads to which users can bring their own tools, as well as submission of change suggestions, analyses, and limited-support projects. Among the latter are a suite of augmented reality tools to visualize the Project data, co-developed by Columbia's Computer Graphic and User Interfaces Lab. At the heart of this Community Crossroads lies Performant Software's release of a community version of the edition software for adaptation to new texts and projects.

NAME: Tianna Helena Uchacz
AFFILIATION: Texas A&M University
EMAIL: tianna.uchacz@gmail.com

TITLE: Collaborative Digital Humanities Teaching as a Research Method: A Case Study toward a Research and Teaching Companion

ABSTRACT: The Making and Knowing Project practices pedagogy-driven research-a method in which students with no necessary prior training undertake intensive skill-building and then consolidate their newly won competencies through credited work on research tasks and problem-based projects that advance the Project's overall research goals. Two new DH graduate seminars at Columbia, co-designed and collaboratively taught in 2017 and 2019, yielded edition prototypes and exploratory textual analyses that proved instrumental in the development of the Project's Digital Critical Edition. This talk presents insights from the design and outcomes of these seminars-insights which stand as one component part of the Project's Research and Teaching Companion (full release expected July 2021), which will offer a formal articulation of the Project's principles and methods of pedagogy-driven research as well as a resource set for other scholars, instructors, and practitioners to scale and adapt to their own particular needs.

NAME: Tillmann Taape
AFFILIATION: Columbia University
EMAIL: tillmann.taape@cantab.net

TITLE: How to Transcribe and Translate an Early Modern How-to Text (and how not to)

ABSTRACT: This paper reviews the textual scholarship principles, pedagogical strategies, and platform decisions behind the Making and Knowing Project's five-year, workshop-based effort to transcribe, translate, and encode BnF Ms. Fr. 640, an eclectic collection of practical knowledge in sixteenth-century French whose prose is idiosyncratic, highly technical, and heterogeneous, ranging from terse to rambling. The collaborative and iterative nature of the undertaking, as well as the deliberate entwining of the text-based work with hands-on historical reconstruction of materials and techniques in the laboratory, led to an evolution of the aims and strategies of our transcriptions and translation over time, with implications for workflow from workshop to workshop and for the eventual presentation of the text in the Edition.

NAME: Clément Godbarge
AFFILIATION: Columbia University
EMAIL: clement.godbarge@fulbrightmail.org

TITLE: Le mille-feuille éditorial et ses ramifications théoriques

ABSTRACT: L'édition critique numérique du BnF Ms. Fr. 640 offre aux lecteurs pas moins de cinq versions différentes du texte: le fac-similé (en mode image), la transcription diplomatique (xml), la transcription normalisée (xml), une traduction anglaise (xml) ainsi que le rendu final de ces textes sur écran (html/ css). Cette prolifération de couches offre de nombreux avantages. Elle permet aux éditions critiques de s'ouvrir

à de nouveaux publics tout en invitant le lecteur à explorer les différentes étapes du processus éditorial et de se familiariser avec la source première. Elle rend aussi l'édition plus ouverte et extensible, facilitant à l'avenir la superposition de couches supplémentaires, comme de nouvelles variantes ou traductions du texte. Cette prolifération de couches, cependant, soulève un certain nombre de questions, liées notamment aux conséquences de cette complexification. En effet, comment additionner les couches sans nuire à la cohérence de notre édition? Quelle place donner au choix éditorial dans un modèle qui en favorise la dilution?

Sustainability Strategies for Digital Humanities Systems

Claes Neuefeind
c.neuefeind@uni-koeln.de
Cologne Center for eHumanities, University of Cologne, Germany

Brigitte Mathiak
bmathiak@uni-koeln.de
Data Center for the Humanities, University of Cologne, Germany

Philip Schildkamp
Data Center for the Humanities, University of Cologne, Germany

Unmil Karadkar
unmil.karadkar@uni-graz.at
Centre for Information Modelling, University of Graz, Austria

Johannes Stigler
Centre for Information Modelling, University of Graz, Austria

Elisabeth Steiner
Centre for Information Modelling, University of Graz, Austria

Gunter Vasold
Centre for Information Modelling, University of Graz, Austria

Fabio Tosques
Centre for Information Modelling, University of Graz, Austria

Arianna Ciula
arianna.ciula@kcl.ac.uk
King's Digital Lab, King's College London, United Kingdom

Brian Maher
brian.maher@kcl.ac.uk
King's Digital Lab, King's College London, United Kingdom

Greg Newton
gregster@uvic.ca
Humanities Computing and Media Centre, University of Victoria, Canada

Stewart Arneil
Humanities Computing and Media Centre, University of Victoria, Canada

Martin Holmes
Humanities Computing and Media Centre, University of Victoria, Canada

Keywords: Sustainability, Research Software, Technology Stacks
Topics: Global, English, Contemporary, data, object, and artefact preservation, sustainable procedures, systems, and methods, Computer science, Informatics

Now that the Digital Humanities (DH) are becoming a well-established research field, producing seminal publications in print as well as digital formats, the time for consolidation has come. It is noteworthy that digital tools and methods from the pioneering days of the DH are degrading and some have already vanished. Therefore, it is urgent to take action and to prevent further losses. While the necessity of high quality research data manage

Now that the Digital Humanities (DH) are becoming a well-established research field, producing seminal publications in print as well as digital formats, the time for consolidation has come. It is noteworthy that digital tools and methods from the pioneering days of the DH are degrading and some have already vanished.

Therefore, it is urgent to take action and to prevent further losses. While the necessity of high quality research data management (RDM) is encouraged or even required by funding agencies and there is an increasing awareness for long-term archiving (LTA), when it comes to primary research data, the fact that the DH exhibit a structural deficit regarding maintaining and preserving research software is at the least underestimated.

In this panel, we will focus on infrastructure and institutional support. Beginning with an overview of existing strategies from the DH and beyond, we highlight selected strategies to compare how they are implemented at different institutions in terms of infrastructure, expert knowledge and also funding. We also want to evaluate the extent of institutional support that is needed to successfully sustain and archive DH projects and the software they use. We will discuss currently implemented solutions to maintain and preserve research projects and software, all of which approach the outlined problem from a different angle.

1. Sustainability Strategies in DH and beyond (Brigitte Mathiak, Data Center for the Humanities, University of Cologne)

Sustainability of research software is an important issue for the DH. In our investigation of the "Digital Scholarly Editions" online catalogue, we compared the time stamps of the last seen version on the Internet Archive with the first seen version (Schildkamp & Mathiak, 2019). We discovered that of 466 digital editions, 376 had disappeared (cf. fig. 1). The average life time is 8.5 years, while the half life time is about 6 years. We expect that other DH projects exhibit similar trends. The reasons for the disappearance of these valuable research resources are manifold: diminishing funding, lack of institutional support and, over time, lack of personnel support as researchers switch career paths or research directions. The "Digital Dark Age" (Whitt, 2017) affects not only our digital cultural heritage, but also the born digital outcomes of scholarly labor.

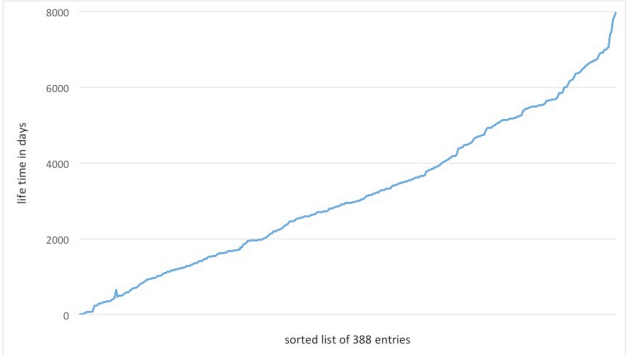


Figure 1: Life time of digital scholarly editions

The problem of sustainability is neither unknown, nor without solutions. Several different models have been explored within the DH community. These include the development of centers such as CHNM, consortia such as Europeana, Hathitrust, and DPLA, as well as community partnerships such as Samvera (previously Hydra) and Islandora. Individual institutions such as those represented on this panel have taken up responsibility for the resources that were placed in their care. Yet, there is a dazzling variety of strategies, technologies, and policies that have been adopted to improve the elusive sustainability, e.g. code archiving, open source dissemination, duplication, sandboxing, refactoring, unified tech stacks, virtual research environments, virtualization, and use of the Internet Archive. While it is clearly easier to prepare a project for sustainability in the planning stage, advice for enhancing sustainability is divergent, ranging from using simple technology, someone's preferred infrastructure, or particular documentation practices. Many completed projects do not have a sustainability strategy, either because they were too old or too optimistic. What happens to these projects is often determined by funding and institutional support. The luxury version is a complete redesign with all the newest bells and whistles, but there are also cheaper strategies, such as putting the system in a sandbox, or relying on the Internet Archive.

However, the problem of sustainability is not unique to the DH. Basic sciences (biology and physics), atmospheric and space sciences, as well as geosciences are some disciplines that are developing sustainability enhancing mechanisms. In conjunction with funding agencies such as the National Science Foundation, researchers in these disciplines have attempted approaches such as community engagement in software and schema development (Specify), ongoing external funding for maintenance (rather than only for new research), long-term funding arcs (NSF centers), funding agency

mandates (contribution of digitized data to existing repositories), efforts to desilo or integrate resources (iDigBio, iPlant), and institutional support for pre-publication drafts (arXiv). We will explore the breadth of these approaches as well as the expected and actual impact of these strategies on sustainability of products that are critical for scholarship in these disciplines, and connect the dots by drawing parallels to the DH.

2. TOSCA-based Application Management (Claes Neuefeind and Philip Schildkamp, Cologne Center for eHumanities/Data Center for the Humanities, University of Cologne)

The University of Cologne's Data Center for the Humanities (DCH) is obliged to concern itself with the sustainability of all digital artifacts produced during (Digital) Humanities projects, e.g. as run by the Cologne Center for eHumanities (CCeH). And as such, it is not only committed to the long-term preservation of data, but of so-called "living systems" (Sahle & Kronenwett, 2013) as well. With regards to this necessity, the DCH is currently engaged in the DFG-funded project "SustainLife - Sustaining Living Digital Systems in the Humanities" (Neuefeind et al., 2018), conducted in cooperation with the Institute of Architecture of Application Systems (IAAS) of the University of Stuttgart. The project aims at adopting the "Topology and Orchestration Specification for Cloud Applications" (TOSCA) standard (OASIS, 2013 and 2019) to the field of Digital Humanities. Being an industry standard focussed on deployment and maintenance of complex software services, TOSCA allows to model applications as abstract topologies consisting of reusable components, while avoiding any kind of vendor or technology lock-in. Through this meta-modelling of software components, not only can the deployment context be adjusted easily (e.g. deployments geared towards OpenStack can easily be adjusted towards Docker, VMWare vSphere, etc.), but from the reusability of said components, synergetic effects emerge, lessening the overall administrative costs for long-term archiving and deployment of research applications.

In our contribution to the panel, we will present the methodological concept of our approach based on the OpenTOSCA ecosystem (Breitenbücher et al., 2016), an open-source implementation of the TOSCA standard, as well as a distinct set of use case implementations conducted within the SustainLife project. The use cases to be presented will cover some of the typical technology stacks in the DH. Foremost, the (1) EarlyCinema use case stands for one of the most common technology stacks: LAMP (Linux, Apache, MySQL, PHP). Further, the (2) AutoPost and (3) TiwoliJ use cases employ the popular Java framework Spring(Boot) with a MySQL database as persistence layer. Also implemented using Spring(Boot), but persisting data in MongoDB, employing ElasticSearch as indexing service and packing a ReactJS frontend, the (4) VedaWeb use case represents one of the more specialized stacks. And lastly the (5) Musical Competitions Database is the most specialized use case, as it depends on older versions of CouchDB for data persistence and ElasticSearch for indexing persisted data (Neuefeind et al., 2019).

3. GAMS: Geisteswissenschaftliches Asset Management System (Unmil Karadkar, Johannes Stigler, Elisabeth Steiner, Gunter Vasold, and Fabio Tosques, Centre for Information Modelling, University of Graz)

Recognizing the problems inherent in conducting digital humanities research based on stand-alone, custom software, The Centre for Information Modeling has developed, maintained, and enriched GAMS - a modular, standards-based, community-used software - since the early 2000's, gaining over 15 years of experience in sustaining a digital scholarship infrastructure. The GAMS infrastructure is supported by ongoing relationships with researchers, personnel, processes, and certifications that inform a holistic, long-term sustainability philosophy. Thus, GAMS embodies a strategy for digital preservation that has been hardened through software upgrades, continuous use, and external testing. GAMS hosts over 95,000 compound digital objects and supports over 90 digital humanities projects.

Infrastructure: The GAMS software is developed using open software and platform-independent standards. These include FEDORA - a flexible open repository infrastructure, Blazegraph, a standards-based, high-performance graph database, Handle - a persistent identifier service, PostgreSQL, Apache Cocoon, Apache Lucene, Apache Solr, and LORIS IIIF image server. GAMS was initially developed using FEDORA 2.0 and over the years, has been migrated to FEDORA 3.5. The GAMS team has developed OAIS-compliant workflows in order to support long-term preservation. Data stored in GAMS is subject to FAIR data principles. Currently, the GAMS team is updating the backend to FEDORA 6.0. This upgrade presents unique challenges as FEDORA has outsourced the notion of content models since version 4.0 and model compliance must now be handled in the application layer. The modularity of the GAMS architecture facilitates such an upgrade as the Java-based Cirilo client supports the management of a legacy layer while migrating to a REST-API-based interface. Cirilo is developed on an open source philosophy and is

available for download via GitHub. The user interface layer of the GAMS Web interface is based on Web technology standards, such as XML and XSLT that separate structure from content and enable multiple, context-specific renditions of Web-based information.

Relationships: In order to ensure continued relevance, the GAMS team partners with humanities researchers. GAMS receives and stores data in recognized archive-compliant standards such as JPEG2000 and TIFF. In addition to interfaces for upload, management, and description of digital objects, the team consults with research partners about issues such as document digitization, ingest, description, and management, developing custom workflows, data models, deposit agreements, data management plans, and publication pipelines as necessary. Developed tools and techniques are available for other projects, thus enriching GAMS as well as the digital research environment for humanists.

Personnel: Continuity of people is often correlated with the availability of infrastructure and data. In order to ensure long-term availability of the data as well as services, the centre invests in project staff for tasks such as software development, infrastructure management, processes, workflows, and content model design, as well as for document and metadata enrichment.

Certifications: In a demonstration of our commitment for long-term preservation and to assure (potential) partners of this commitment, GAMS has undergone rigorous evaluation and has been certified as a trusted digital repository (since 2014), carries the CoreTrustSeal (since 2019), and is registered with the Registry of Research Data Repositories (ROAR).

4. Managing 100 DH legacy projects and building new ones: a pragmatic and holistic approach to archiving and sustainability (Arianna Ciula and Brian Maher, King's Digital Lab, King's College London)

King's Digital Lab's (KDL) (King's College London) contribution to archiving and sustainability practices in Digital Humanities (DH) will be presented along the following dimensions:

(Human) Sustainability of expertise: As generational change occurs and in line with reorientations across the DH community (see Boyles et al., 2018), it has become increasingly clear that the surest way to sustainability is to ensure continuity of technical expertise, domain knowledge, and tacit understanding. KDL conceived and adopted a relatively flexible model with defined career development document and Research Software Engineering (RSE) role definitions (Smithies, 2019).

(Technical) Sustainability of systems and technical stack: The second dimension needed to sustain the DH tradition and fulfil KDL's mandate to increase digital capability across the Arts & Humanities is caring for the cluster of technical systems comprised of hardware and software, web servers, network infrastructure, application frameworks, programming languages, tools (for project, data and code management), and equipment. In practice, sustainable management of lab projects required the adoption of limited server and development environment stacks, in a move away from the more flexible but difficult to manage environment used in earlier eras (for more details of the tools used to support the stack see https://stackshare.io/kings-digital-lab).

(Operational) Post-project integrated in the lab Software Development Lifecycle: The techniques used to manage KDL rich and heterogeneous estate of legacy projects matured into an ongoing process of archiving and sustainability tailored to the Lab's historical, technical and business context. It is applied to new as well as legacy projects, in a manner that ensure systems as well as data are maintained throughout defined life-cycles (King's Digital Lab, 2019). To control this, open ended Service Level Agreement (SLA) contracts are offered to Principal Investigators (PIs) of collaborative research projects to secure maintenance of legacy projects in their live state; however, other options for archiving are also possible and assessed (see also Smithies et al. 2019 and https://dev.clariah.nl/files/dh2019/boa/0648.html). To make the overall approach sustainable, it had to be integrated into the lab's Software Development Lifecycle (SDLC; see https://kingsdigitallab.github.io/sdlc-for-rse/), and in so doing align with KDL infrastructure architecture and core technical stack, while at the same time informing practices of forward planning for new projects. KDL's contribution to the panel will reflect on how alignment across these three layers raises challenges but also poses the foundations for the sustainability of the Lab's ecosystem, hopefully offering a reference for others to reflect upon, adapt and improve.

5. Keeping it Simple and Straightforward (Greg Newton, Stewart Arneil, Martin Holmes, Humanities Computing and Media Centre, University of Victoria)

The University of Victoria long ago demonstrated its commitment to DH research by providing base-budget funding for the five-person Humanities Computing and Media Centre (see https://www.uvic.ca/humanities/hcmc/) - a department in the Faculty of Humanities. As can be seen from the name, HCMC actually pre-dates the term Digital Humanities. As a base-budget funded department, HCMC has the capacity to take on projects regardless of their level of funding - we regularly take on projects with no funding at all - and the commitment to support the project's outputs in perpetuity. This is

only possible due to a critical mass of professors and executives seeing value over time.

For over twenty years HCMC has been consulting on and developing web applications in support of teaching and research. On behalf of our academic collaborators we work closely with library and systems colleagues who take primary responsibility for archiving and technical infrastructure, respectively. This is a strategic division of labour entailing ongoing communications with the benefits of specialization and scale.

Over the years we have come to recognize the inherent dangers of creating teetering stacks of complicated, fashionable technology that cannot stand the test of time. Experiments with several CMS's, JavaScript libraries, and so forth has invariably led us to the conclusion that the long-term cost of coping with breakage and security problems outweighs the short-term value these applications and libraries offer.

While we provide institutional support we are not keen on a never-ending cycle of upgrades and code-maintenance. To mitigate this we have become staunch supporters of KISS - in our case it might stand for "Keep It Simple and Straightforward". We take on very few projects that we did not develop, and when we do they are usually converted to a static site and archived or completely re-written.

Our Project Endings (see https://projectendings.github.io/) survey and interviews have made us doubly aware of the potential for catastrophe using technology that is not proven to be simple and durable. From our perspective every project will benefit from adopting a KISS strategy, but perhaps especially those projects without institutional support.

References

Arneil, Stewart, Martin Holmes and Greg Newton. 2019. "Project Endings: Early Impressions From Our Recent Survey On Project Longevity In DH." Digital Humanities 2019 Conference, Utrecht, Netherlands. July 10 2019. https://dev.clariah.nl/files/dh2019/boa/0891.html

Boyles, Christina, Carrie Johnston, Jim McGrath, Paige Morgan, Miriam Posner, and Chelcie Rowell. 2018. 'Precarious Labor in the Digital Humanities - DH 2018'. In Digital Humanities 2018: Book of Abstracts / Libro de Resúmenes, edited by Élika Ortega, Glen Worthey, Isabel Galina, and Ernesto Priani, 47-52. Mexico City: Red de Humanidades Digitales A.C. https://dh2018.adho.org/precarious-labor-in-the-digital-humanities/.

Breitenbücher, Uwe, Endres, C., Képes, K., Kopp, O., Leymann, F., Wagner, S., Wettinger, J., Zimmermann, M. 2016. 'The OpenTOSCA Ecosystem. Concept & Tools'. In: European Space project on Smart Systems, Big Data, Future Internet - Towards Serving the Grand Societal Challenges - Volume 1: EPS Rome 2016. SciTePress, pp. 112-130.

King's Digital Lab. 2019. 'Archiving and Sustainability | King's Digital Lab'. King's Digital Lab. 2019. https://www.kdl.kcl.ac.uk/our-work/archiving-sustainability/.

Neuefeind, Claes, Lukas Harzenetter, Philip Schildkamp, Uwe Breitenbücher, Brigitte Mathiak, Johanna Barzen, Frank Leymann. 2018. 'The SustainLife Project - Living Systems in Digital Humanities'. In: Proceedings of the 12th Advanced Summer School on Service-Oriented Computing (SummerSoC 2018) (IBM Research Report RC25681), pp. 101-112.

Neuefeind, C. and Schildkamp, P. and Mathiak, B. and Marčić, A. and Hentschel, F. and Harzenetter, L. and Breitenbücher, U. and Barzen, J. and Leymann, F. (2019).

Sustaining the Musical Competitions Database. A TOSCA-based Approach to Application Preservation in the Digital Humanities. In: Book of Abstracts of the 29th Digital Humanities Conference (DH 2019), https://dev.clariah.nl/files/dh2019/boa/0574.html

OASIS. 2013. Topology and Orchestration Specification for Cloud Applications Versio 1.0, http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.html

OASIS. 2019. TOSCA Simple Profile in YAML Version 1.2, http://docs.oasisopen. org/tosca/TOSCA-Simple-Profile-YAML/v1.2/TOSCA-Simple-Profile-YAMLv1.2.html.

Sahle, Patrick and Simone Kronenwett. 2013. 'Jenseits der Daten. Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner Dat a Center for the Humanities'. In: LIBREAS. Library Ideas #23, pp. 76-96.

Schildkamp, Philip and Brigitte Mathiak. 2019. 'Overview of life and death of digital scholarly editions' [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3483998

Smithies, James. 2019. 'The Continuum Approach to Career Development: Research Software Careers in King's Digital Lab'. King's Digital Lab - Thoughts and Reflections from the Lab (blog). 7 February 2019. https://www.kdl.kcl.ac.uk/blog/rse-careerdevelopment/.

Smithies, James, Anna Maria Sichani, Carina Westling, Pam Mellen, and Arianna Ciula. 2019. 'Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab'. Digital Humanities Quarterly. http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html.

Whitt, Richard. 2017. 'Through A Glass, Darkly' Technical, Policy, and Financial Actions to Avert the Coming Digital Dark Ages, 33 Santa

Clara High Tech. L.J. 117. Available at: http://digitalcommons.law.scu.edu/chtlj/vol33/iss2/1

ment (RDM) is encouraged or even required by funding agencies and there is an increasing awareness for long-term archiving (LTA), when it comes to primary research data, the fact that the DH exhibit a structural deficit regarding maintaining and preserving research software is at the least underestimated.

In this panel, we will focus on infrastructure and institutional support. Beginning with an overview of existing strategies from the DH and beyond, we highlight selected strategies to compare how they are implemented at different institutions in terms of infrastructure, expert knowledge and also funding. We also want to evaluate the extent of institutional support that is needed to successfully sustain and archive DH projects and the software they use. We will discuss currently implemented solutions to maintain and preserve research projects and software, all of which approach the outlined problem from a different angle.

1. Sustainability Strategies in DH and beyond (Brigitte Mathiak, Data Center for the Humanities, University of Cologne)

Sustainability of research software is an important issue for the DH. In our investigation of the "Digital Scholarly Editions" online catalogue, we compared the time stamps of the last seen version on the Internet Archive with the first seen version (Schildkamp & Mathiak, 2019). We discovered that of 466 digital editions, 376 had disappeared. The average life time is 8.5 years, while the half life time is about 6 years. We expect that other DH projects exhibit similar trends. The reasons for the disappearance of these valuable research resources are manifold: diminishing funding, lack of institutional support and, over time, lack of personnel support as researchers switch career paths or research directions. The "Digital Dark Age" (Whitt, 2017) affects not only our digital cultural heritage, but also the born digital outcomes of scholarly labor.

The problem of sustainability is neither unknown, nor without solutions. Several different models have been explored within the DH community. These include the development of centers such as CHNM, consortia such as Europeana, Hathitrust, and DPLA, as well as community partnerships such as Samvera (previously Hydra) and Islandora. Individual institutions such as those represented on this panel have taken up responsibility for the resources that were placed in their care. Yet, there is a dazzling variety of strategies, technologies, and policies that have been adopted to improve the elusive sustainability, e.g. code archiving, open source dissemination, duplication, sandboxing, refactoring, unified tech stacks, virtual research environments, virtualization, and use of the Internet Archive. While it is clearly easier to prepare a project for sustainability in the planning stage, advice for enhancing sustainability is divergent, ranging from using simple technology, someone's preferred infrastructure, or particular documentation practices. Many completed projects do not have a sustainability strategy, either because they were too old or too optimistic. What happens to these projects is often determined by funding and institutional support. The luxury version is a complete redesign with all the newest bells and whistles, but there are also cheaper strategies, such as putting the system in a sandbox, or relying on the Internet Archive.

However, the problem of sustainability is not unique to the DH. Basic sciences (biology and physics), atmospheric and space sciences, as well as geosciences are some disciplines that are developing sustainability enhancing mechanisms. In conjunction with funding agencies such as the National Science Foundation, researchers in these disciplines have attempted approaches such as community engagement in software and schema development (Specify), ongoing external funding for maintenance (rather than only for new research), long-term funding arcs (NSF centers), funding agency mandates (contribution of digitized data to existing repositories), efforts to desilo or integrate resources (iDigBio, iPlant), and institutional support for pre-publication drafts (arXiv). We will explore the breadth of these approaches as well as the expected and actual impact of these strategies on sustainability of products that are critical for scholarship in these disciplines, and connect the dots by drawing parallels to the DH.

2. TOSCA-based Application Management (Claes Neuefeind and Philip Schildkamp, Cologne Center for eHumanities/Data Center for the Humanities, University of Cologne)

The University of Cologne's Data Center for the Humanities (DCH) is obliged to concern itself with the sustainability of all digital artifacts produced during (Digital) Humanities projects, e.g. as run by the Cologne Center for eHumanities (CCeH). And as such, it is not only committed to the long-term preservation of data, but of so-called "living systems" (Sahle & Kronenwett, 2013) as well. With regards to this necessity, the DCH is currently engaged in the DFG-funded project "SustainLife - Sustaining Living Digital Systems in the Humanities" (Neuefeind et al., 2018), conducted in cooperation with the Institute of Architecture of Application Systems (IAAS) of the University of Stuttgart. The project aims at adopting

the "Topology and Orchestration Specification for Cloud Applications" (TOSCA) standard (OASIS, 2013 and 2019) to the field of Digital Humanities. Being an industry standard focussed on deployment and maintenance of complex software services, TOSCA allows to model applications as abstract topologies consisting of reusable components, while avoiding any kind of vendor or technology lock-in. Through this meta-modelling of software components, not only can the deployment context be adjusted easily (e.g. deployments geared towards OpenStack can easily be adjusted towards Docker, VMWare vSphere, etc.), but from the reusability of said components, synergetic effects emerge, lessening the overall administrative costs for long-term archiving and deployment of research applications.

In our contribution to the panel, we will present the methodological concept of our approach based on the OpenTOSCA ecosystem (Breitenbücher et al., 2016), an open-source implementation of the TOSCA standard, as well as a distinct set of use case implementations conducted within the SustainLife project. The use cases to be presented will cover some of the typical technology stacks in the DH. Foremost, the (1) EarlyCinema use case stands for one of the most common technology stacks: LAMP (Linux, Apache, MySQL, PHP). Further, the (2) AutoPost and (3) TiwoliJ use cases employ the popular Java framework Spring(Boot) with a MySQL database as persistence layer. Also implemented using Spring(Boot), but persisting data in MongoDB, employing ElasticSearch as indexing service and packing a ReactJS frontend, the (4) VedaWeb use case represents one of the more specialized stacks. And lastly the (5) Musical Competitions Database is the most specialized use case, as it depends on older versions of CouchDB for data persistence and ElasticSearch for indexing persisted data (Neuefeind et al., 2019).

3. GAMS: Geisteswissenschaftliches Asset Management System (Unmil Karadkar, Johannes Stigler, Elisabeth Steiner, Gunter Vasold, and Fabio Tosques, Centre for Information Modelling, University of Graz)

Recognizing the problems inherent in conducting digital humanities research based on stand-alone, custom software, The Centre for Information Modeling has developed, maintained, and enriched GAMS - a modular, standards-based, community-used software - since the early 2000's, gaining over 15 years of experience in sustaining a digital scholarship infrastructure. The GAMS infrastructure is supported by ongoing relationships with researchers, personnel, processes, and certifications that inform a holistic, long-term sustainability philosophy. Thus, GAMS embodies a strategy for digital preservation that has been hardened through software upgrades, continuous use, and external testing. GAMS hosts over 95,000 compound digital objects and supports over 90 digital humanities projects.

Infrastructure: The GAMS software is developed using open software and platform-independent standards. These include FEDORA - a flexible open repository infrastructure, Blazegraph, a standards-based, high-performance graph database, Handle - a persistent identifier service, PostgreSQL, Apache Cocoon, Apache Lucene, Apache Solr, and LORIS IIIF image server. GAMS was initially developed using FEDORA 2.0 and over the years, has been migrated to FEDORA 3.5. The GAMS team has developed OAIS-compliant workflows in order to support long-term preservation. Data stored in GAMS is subject to FAIR data principles. Currently, the GAMS team is updating the backend to FEDORA 6.0. This upgrade presents unique challenges as FEDORA has outsourced the notion of content models since version 4.0 and model compliance must now be handled in the application layer. The modularity of the GAMS architecture facilitates such an upgrade as the Java-based Cirilo client supports the management of a legacy layer while migrating to a REST-API-based interface. Cirilo is developed on an open source philosophy and is available for download via GitHub. The user interface layer of the GAMS Web interface is based on Web technology standards, such as XML and XSLT that separate structure from content and enable multiple, context-specific renditions of Web-based information.

Relationships: In order to ensure continued relevance, the GAMS team partners with humanities researchers. GAMS receives and stores data in recognized archive-compliant standards such as JPEG2000 and TIFF. In addition to interfaces for upload, management, and description of digital objects, the team consults with research partners about issues such as document digitization, ingest, description, and management, developing custom workflows, data models, deposit agreements, data management plans, and publication pipelines as necessary. Developed tools and techniques are available for other projects, thus enriching GAMS as well as the digital research environment for humanists.

Personnel: Continuity of people is often correlated with the availability of infrastructure and data. In order to ensure long-term availability of the data as well as services, the centre invests in project staff for tasks such as software development, infrastructure management, processes, workflows, and content model

design, as well as for document and metadata enrichment.

Certifications: In a demonstration of our commitment for long-term preservation and to assure (potential) partners of this commitment, GAMS has undergone rigorous evaluation and has been certified as a trusted digital repository (since 2014), carries the CoreTrustSeal (since 2019), and is registered with the Registry of Research Data Repositories (ROAR).

4. Managing 100 DH legacy projects and building new ones: a pragmatic and holistic approach to archiving and sustainability (Arianna Ciula and Brian Maher, King's Digital Lab, King's College London)

King's Digital Lab's (KDL) (King's College London) contribution to archiving and sustainability practices in Digital Humanities (DH) will be presented along the following dimensions:

(Human) Sustainability of expertise: As generational change occurs and in line with reorientations across the DH community (see Boyles et al., 2018), it has become increasingly clear that the surest way to sustainability is to ensure continuit

#whereislatinxdh

Gabriela Baeza Ventura
gbventura@uh.edu
University of Houston, United States of America

María Cotera
mcotera@umich.edu
University of Michigan

Linda García Merchant
garciamerchant@gmail.com
University of Nebraska

Marco Seiferle-Valencia
marcomarias@gmail.com
University of Idaho

KarenMary Davalos
kdavalos@umn.edu
University of Minnesota

AnneMarie Pérez
aperez@csudh.edu
CSU-Dominguez Hills

Keywords: Latinx, Archives, Digital archive, Chicanx, Spanish
Topics: English, Spanish, North America, 19th Century, 20th Century, Contemporary, digital activism and advocacy, digital archiving, Chicano/a/x, Latino/a/x studies, Gender and sexuality studies

Through movements, organizations and pedagogical initiatives, such as transformDH, FemTechNet and DOOCs, the field of digital humanities was invited to rethink its initial traditional and anglocentric structures. Although the field has made efforts to center people and knowledges not included in its initial work, there is still much to be done.

On this panel, each Latinx presenter will speak about their work in DH, discuss how their projects/work resists traditional anglocentric/Eurocentric/phallocentric epistemologies, the methodologies they use to create knowledge in order to fight the erasure of Latinx voices, and address how a radical Latinx feminist perspective can work in conjunction with the philosophy of open education to create curriculums using primary archives and student and community knowledge to trouble the standard narrative in the process of creating Openly licensed resources (OER).

The Web in Amber: How the Ephemeral Human Digital Experience is Preserved and Analyzed

Olga Holownia
Olga.Holownia@bl.uk
IIPC & British Library

Sally Chambers
Ghent Centre for Digital Humanities, University of Ghent

Grace Thomas
grth@loc.gov
Library of Congress

Ian Milligan
i2milligan@uwaterloo.ca
University of Waterloo

Ed Summers
edsu@umd.edu
Maryland Institute for Technology in the Humanities

Bergis Jules
Documenting the Now

Mark Phillips
mark.phillips@unt.edu
University of North Texas Libraries

Keywords: web archiving, collections as data, research use, social media, machine learning
Topics: Global, English, 20th Century, Contemporary, data, object, and artefact preservation, digital archiving, Humanities computing, Library & information science

For more than two decades, libraries, universities, archives, non-profits, and other cultural heritage institutions have been archiving the evolution of human society, government, and culture now ephemerally published on the Web. Through domain harvests, collections developed around themes or events, or simply necessity of institutional record, these organizations have built archives of enormous size and richness.

The aim of our panel is to demonstrate how web archives are being used in digital humanities research, including novel approaches to access, analytical tools, machine learning and social media archiving. We also show the importance of community building, collaborations with heritage institutions and offering web archives as one of many resources available to DH researchers - all of which helps bridge the gap between potential and use.

Deciphering Encoded Music: Increasing Access to Open Data Resources

Joanna Devaney
johanna.devaney@brooklyn.cuny.edu
Brooklyn College

Natalia Ermolaev
nataliae@princeton.edu
Princeton University

Anna Kijas
anna.kijas@tufts.edu
Tufts University

Anna Neovesky
Anna.Neovesky@adwmainz.de
ADW Mainz

Mark Saccomano
mss2221@columbia.edu
Columbia University

Raffaele Viglianti
rviglian@umd.edu
Maryland Institute for Technology in the Humanities

Frederic von Vlahovits
Frederic.vonVlahovits@adwmainz.de
ADW Mainz

Keywords: music, accessibility, retrieval, annotation, performance
Topics: Europe, English, North America, 5th-14th Century, 15th-17th Century, 20th Century, digital archiving, music and sound digitization, encoding, and analysis, Humanities computing, Musicology

This panel brings together developers of digital music projects to articulate a vision for the future of open data with regard to the needs and desires of researchers in music studies. How can individual projects contribute to a network of open data sources? What are the obstacles ahead and how can they be overcome? Panel participants will approach these questions from various angles, through discussion of topics that include: metasearch capabilities to retrieve data from multiple repositories of encoded music; increasing data usability by providing multiple access points; developing a critical pedagogy that challenges the social structures underlying standardized encoding schemes; allowing encoded music annotation across repositories and formats; and the development of minimal computing models to enhance possibilities for discoverability and display of individual corpora for analysis.

Unmaking/Remaking Memory Work: Centering Community Narratives of Latinx Lived Experience

Marisa Hicks-Alcaraz
mhicksalcaraz@gmail.com
Claremont Graduate University

William Camargo
william.camargo@cgu.edu
Claremont Graduate University

Joseph Valencia
joedvalencia@gmail.com
University of Southern California

Samantha Cabrera Friend
samanthamariafriend@gmail.com
The Visual Desk School of Visual Arts in Manhattan

Keywords: community, archiving, access, Latinx, accountability
Topics: English, North America, Contemporary, digital activism and advocacy, public humanities collaborations and methods, Chicano/a/x, Latino/a/x studies, Media studies

Panel Abstract:

This panel locates itself within a growing social movement of new community-centered archiving and curatorial initiatives that has risen in recognition that traditionally ignored communities should have a role in how their histories get told. Our papers showcase a range of projects and vantage points aimed at giving visibility to Latinx lived experience, while taking steps to radically unmake/remake archival practices and memory work in a way that is accountable to and in participation with Latinx communities. Our projects are grounded in justice and community-oriented work spanning disparate academic and artistic fields, including photography, critical curatorial and archival praxis, visual studies, and digital humanities. We ask and then demonstrate: What might be alternative means for recovering hidden stories of the Latinx diaspora? How do we practice memory work in a way that is accountable to and in participation with its subjects? What might be new methods for archival and curatorial work that are in accordance with the lived realities of the communities we are working with? We invite discussion to help further develop community-centered initiatives and hope that audience members will leave with new ideas for community-engaged and justice-based work.

Paper Abstracts:

Marisa Hicks-Alcaraz, PhD Student in Cultural Studies at Claremont Graduate University and co-founder of ImaginX en Movimiento

"ImaginX en Movimiento: Building a Rasquache Digital Archive"

ImaginX en Movimiento (IXeM) is a Los Angeles-based moving image archive and collective I co-founded (with one foot in and one foot out of academia) in May 2019 in effort to counter exclusionary practices that prevent community access and participation in cultural memory work. This paper is an ethnographic case study that provides critical reflections on the strategies employed by IXeM that respond to particular constraints and challenges pertaining to labor and cost, including operations, acquisition of equipment, funding, and collaborative initiatives for community engagement. This case study will posit labor and cost solutions against a background of scarce resources as well as forge new ideas for community-engaged and justice-based recuperation methods in academia.

In particular, this paper reflects on the way IXeM expresses a rasquache defiance that is situated within digital ethnic studies and the larger movement of community-based archiving, as well as in the interrogation of traditional (Anglo, heteropatriarchal-centric) archival authority. According to Chicanx art historian, Tomás Ybarra-Frausto, being rasquache is the visceral response to lived reality and an attitude rooted in resourcefulness and adaptability (1989). The Chicanx phenomenon of rasquachismo, or the view of the downtrodden, encourages its practitioners to use the materials available at hand to create something that is accessible and belongs to the people. A deliberate choice, rasquachismo is political and subversive; a tool of activism and an anti-elite practice. In this paper, I develop the idea of digital rasquachismo as a way to dismantle traditional archival practices that reinforce oppressive systems that disproportionately subject Brown and Black bodies to generational inequities and inspire alternative methods that retool digital technologies to center the lived experiences and wellbeing of those at the margins.

Blurring the divide between choice and necessity, IXeM is driven by both a socioeconomic imperative and a deliberate choice to challenge the authority of top-down knowledge systems by retooling commercial

content sharing platforms as archival databases (see: @veteranas_and_rucas, @blvckvrchives, @latinx_diaspora_archives, @quinceaneraarchives, LA Freewaves), which some scholars have viewed as revolting and others an avenue of revolt. Going beyond just "making do," IXeM enacts a radical politics of resistance by grabbing hold of the tools of social media to transform cultural memory work and its publics; in effect, using the master's tools to dismantle the master's house.

While we see such platforms as presenting opportunities to transform the practice of social memory, we also firmly acknowledge the ways in which they can be (and have been) used to cause harm and facilitate surveillance and oppression. As such, this talk will consider both the ethical risks and opportunities afforded by these new tools, especially as they pertain to documenting Latinx grief and trauma in the wake of the on-going COVID-19 pandemic, its economic fallout, and the current uprisings against racist police brutality.

Bibliography:

Bray, Anne. LA Freewaves. http://freewaves.org/

Cherlise, Renata. @blvckvrchives. https://www.instagram.com/blvckvrchives/

Camargo, William. @latinx_diaspora_archives. https://www.instagram.com/latinx_diaspora_archives/

Cabrera Friend, Samantha. @quinceaneraarchives. https://www.instagram.com/quinceaneraarchives/

Rosales, Guadalupe. @veteranas_and_rucas. http://architectureofnecessity.com/

Ybarra-Frausto, Tomás. "Rasquachismo: a Chicano sensibility." Chicano Aesthetics: Rasquachismo, edited by Rudy Guglielmo et al., MARS, 1989.

Joseph Daniel Valencia, Independent Curator

"Curating Together: Decolonial and Polyvocal Approaches to Exhibition-Making"

This presentation seeks to radically reimagine curatorial practice to position the curator not as an authoritative expert, but instead as an organizer and collaborator with community members located inside and outside of the institution. It is structured around the following question: What happens when one curates within communities of color, particularly Latinx populations, and translates these community narratives into an institutional context?

Through the case study of the exhibition "Liberate the Bar! Queer Nightlife, Activism, and Spacemaking" at the ONE Gallery in West Hollywood, CA (2019), I reflect on the function of community-centered curating in altering the narrative, structure, ownership, and impact of the exhibition. I examine how the practice of "curating together" is inherently decolonial, counter-hegemonic, and polyvocal. I also consider how this practice can expand the parameters of an exhibition, thinking through the ways community knowledge-sharing, public programs, and digital initiatives can be understood as an extension of a curatorial project. The ultimate goal of this presentation is to explore how curating can be practiced and understood as it relates to diverse populations and their lived experiences.

William Camargo: Visual Artist/Art Educator, founder of Latinx Diaspora Archives

"@Latinx_Diaspora_Archives: Digital Archives a Wave of Digital Activism On the Rise"

The Instagram account I created, @latinx_diaspora_archives, is an attempt to elevate communities of color, those specifically of the Latinx diaspora. While this is a huge task for anyone to undertake, many have succeeded. This paper looks at various accounts that have been started by other non-traditional archivists, such as @atx_barrio_archives, @veteranasyrucas, @fadedresistance @archivotrans and many more as a new movement of the archive as activism. It is also an attempt to question the institution of the archive: Who holds the majority of archives of black folks? Of queer folks? Of Latinx communities and other marginalized groups? Who has access to them? How are they displayed? Coming from a low-income background, an attempt to look at the archives posed difficulties, the resources and access were not there. @Latinx_Diaspora_Archives attempts to center communities of color as part of the "American" experience. It aims to center family photos of the Latinx Diaspora through the Instagram account in order to combat erasure throughout the U.S. and the Americas. All of the Instagram accounts mentioned here challenge hegemonic and hierarchical systems that have been set up by institutions.

I have openly discussed that I myself am a non-traditional archivist. The ownership of each family photograph that I receive remains with the family and is returned to the owner(s), unless gifted to me. It is an attempt to give a voice to the narratives of the community. In Jose Esteban Munoz's "Ephemera as Evidence: Introductory Notes to Queer Acts," he points to who is left out of the archive in the traditional sense: queer, communities of color, and other marginal communities. The increase of Instagram accounts lets us into these ephemeral moments in these communities, through the eyes of those that experienced them. We have seen when holders of archives and photographers, who either photograph or distribute photographic archives, further push negative stereotypes about our communities. This project is an attempt to elevate our communities and show our realities, through our own experiences and through the radical archives we are a part of.

Samantha Cabrera Friend, Visual Artist and founder of Quinceañera Archives

"Consumerism in Digital Archives: Placing Latinx Traditions on the Right Cultural Shelf"

To assign attribution and credit fairly" is one of the hallmarks of many digital humanities, professional archiving and open data practices. But when we view the practices in place to acknowledge any initial cultural significance of modern findings passed over by the contributors themselves, we run into a murkier, much more complex procedure.

Take any household product which reads "MADE IN CHINA" on the body of the item itself and again on the packaging. Do we feel the factual essence of its Chinese origins in our possession, or out of necessity, convenience, and immersion are we utilizing that tool i n a completely new, aptly disregarding space?

This simplified example can be most intricately applied to the case of the modern quinceañera, an indigenous rite of passage from 1400s México that has (in the most accessible instances) been viewed through a North American lens. One may inherently acknowledge the tradition's roots by use of the Spanish word quinceañera, but the idea as commonly keyworded, categorized and understood in the United States (as comparable to the Sweet 16 among other celebratory practices) remains misguided.

Latinx customs, rituals and traditions have withstood countless transformations through their assimilation to the US and back, and the undocumented history of the quinceañera-a journey I hold quite close to me-is still just one of the many deprived of proper accreditation. So was born, Quinceañera Archives, an online community by way of Instagram. Consequently, this proposal explores the concept that in light of the Internet, Latinx traditions, communities and cultures, have reached increasing new levels of public visibility yes, but at the frequent expense of accessibility, by nature of false ownership, misinterpretation, and selective distribution. Do we work towards collaboration with larger institutions in hopes of maximum visibility or place our effort in grassroots movements which need no reclassification or permittance, in order to ensure accessibility? The benefits of both sides are those I seek to highlight through a multiplicity of methods which I study in quinceañera practice-working to place Latinx narrative histories on the "right" cultural shelf. Though success in relaying this message is never guaranteed, there is great comfort in knowing that the work is now in our hands.

Operationalizing Surveillance Studies in the Digital Humanities

Christina Boyles
cboyles@msu.edu
Michigan State University, United States of America

Arun Jacob
arun.jacob@mail.utoronto.ca
University of Toronto, Canada

Andy Boyles Petersen
andyjp@msu.edu
Michigan State University, United States of America

Megan Wilson
megeileenwilson@gmail.com
Independent Scholar, Canada

Keywords: surveillance, ethics, methodologies, technology, intersectionality
Topics: English, North America, Contemporary, digital access, privacy, and ethics analysis, meta-criticism (reflections on digital humanities and humanities computing), Cultural studies, Media studies

For critical digital humanists, surveillance is a site where issues surrounding race, gender, and sexuality intersect with our digital lives. Works such as Simone Browne's Dark Matters: On the Surveillance of Blackness reveal a growing concern with surveillance, with groups such as SurvDH and the Digital Library Federation's Technologies of Surveillance (ToS) working group providing outlets for digital humanists to explore these topics in more depth.

Digital humanists are searching for a clear articulation of our field's engagement with surveillance theory/studies. This panel brings together the founders of SurvDH to develop a praxis of (counter)surveillance in the digital humanities. In particular, we will discuss:

- Rising applications of surveillance technologies in digital scholarship, including algorithmic design, Ed Tech, and facial recognition software.
- Resisting techno-solutionism in both project development and theorization
- Operationalizing surveillance studies methodologies in DH research

- Our civic and ethical responsibility as digital humanists

Markup as Critical Pedagogy

Elisa Eileen Beshero-Bondar
ebb8@pitt.edu
University of Pittsburgh at Greensburg, United States of America

Janelle Jenstad
jenstad@uvic.ca
University of Victoria, Canada

Jessica H. Lu
jhl.jessica@gmail.com
University of Maryland, College Park

Paul Schacht
schacht@geneseo.edu
SUNY Geneseo

Rebecca Nesvet
nesvetr@uwgb.edu
University of Wisconsin Green Bay

Keywords: TEI, XML, critical pedagogy, code switching, XSLT
Topics: English, North America, Contemporary, scholarly editing and editions development, analysis, and methods, text encoding and markup language creation, deployment, and analysis, Education/ pedagogy, Humanities computing

This panel addresses the teaching of markup and the XML family of languages as a critical pedagogy that empowers students as decision makers and project managers. The panelists discuss how the exploration, schematization, documentation, transformation, and review of markup interpellates newcomers to digital humanities as thoughtful participants in real textual communities. We find that the practice of sharing markup languages with students counters persistent misconceptions in the DH community about hierarchical data modeling with XML and TEI. Rather than experiencing conceptual limitation in the tree structure, students practicing markup and its transformations engage in a direct confrontation with cultural and computational issues surrounding the crafting, control, and distribution of documents--a confrontation that involves code switching and intervention in controlling discourses.

Ethical Soundings in Collaborative Digital Humanities Research Projects: Critical Scenarios from The SpokenWeb

Jason Camlot
jason.camlot@concordia.ca
Concordia University, Canada

Tanya Clement
tclement@utexas.edu
University of Texas at Austin

Klara du Plessis
klaraduplessis@gmail.com
Concordia University, Canada

Liz Fisher
lfisher@austin.utexas.edu
University of Texas at Austin

Deanna Fong
deanna.m.fong@gmail.com
Concordia University, Canada

Yuliya Kondratenko
yuliya.kondratenko@concordia.ca
Concordia University, Canada

Emily Murphy
emurph01@mail.ubc.ca
University of British Columbia, Okanagan

Annie Murray
amurr@ucalgary.ca
University of Alberta

Michael O'Driscoll
mo@ualberta.ca
University of Calgary

Karis Shearer
karis.shearer@ubc.ca
University of British Columbia, Okanagan

Keywords: Ethics, Collaboration, Research Networks, Literature, Sound Studies
Topics: EnglishNorth America20th Century, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), public humanities collaborations and methods, Cultural studies, Literary studies

Ethical Soundings in Collaborative Digital Humanities Research Projects: Critical Scenarios from The SpokenWeb

Authors:

(1) Jason Camlot, Professor and Research Chair in Literature and Sound Studies, Concordia University;

(2) Tanya Clement, Associate Professor, Digital Humanities and English, University of Texas;

(3) Klara du Plessis, Doctoral Candidate, English, Concordia University;

(4) Liz Fisher, Doctoral Candidate, Digital Humanities and English, University of Texas;

(5) Deanna Fong, Postdoctoral Fellow in Literature and Oral History, Concordia University;

(6) Yuliya Kondratenko, SpokenWeb Project Manager, Concordia University;

(7) Emily Murphy, Assistant Professor, Digital Humanities, Department of English and Cultural Studies, University of British Columbia, Okanagan;

(8) Annie Murray, Associate University Librarian for Archives and Special Collections, University of Calgary;

(9) Michael, O'Driscoll, Professor, English and Film Studies, Vice-Dean of Arts, University of Alberta;

(10) Karis Shearer, Associate Professor and Principal's Research Chair, English and Cultural Studies, University of British Columbia, Okanagan.

This series of papers draws upon recent and ongoing experiences from the interdisciplinary SpokenWeb research programme to consider the range of ethical scenarios that inform this collaborative network's engagement with audio archives of literary and humanities-oriented sound recordings, and the communities of practice that generated them. We will consider ethical scenarios through specific projects of the SpokenWeb partnership, a research network that aims to develop coordinated and collaborative approaches to literary historical study, digital development, and critical and pedagogical engagement with diverse collections of literary sound recordings from across Canada and beyond. The goals and projects of the SpokenWeb partnership that will serve to focus our discussion of ethical scenarios in large-scale collaborative digital humanities networks include:

1. rights and access management of digital research data and metadata;

2. the ethics of archival listening as they pertain to the development of new forms of historical

and critical scholarly methods of engagement with the contents of documentary audio

Archives;

3. automated techniques and tools for searching, visualizing, analyzing and enhancing critical

engagement (for features relevant to humanities research and pedagogy);

4. pedagogy, training, mentorship and student labor (the organization of roles and relations

across the research network);

5. innovative ways of mobilizing digitized spoken and literary recordings within performative and

public contexts;

6. project management and governance.

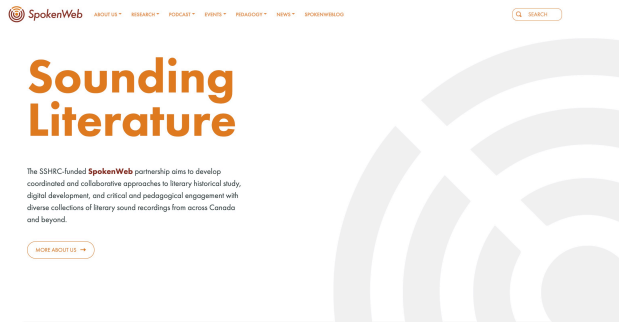


Figure 1. Mainpage of SpokenWeb Research Site

To date, the proposed SpokenWeb archive contains approximately 5000 hours, or 12Tb, of audio data. The SpokenWeb program begins with the preservation and description of sonic artifacts that have captured literary events of the past, and quickly moves into a wide range of approaches and activities activate these artifacts in the present. For the purpose of this set of presentations that will use SpokenWeb projects as case studies for the consideration of ethical scenarios in digital humanities research programs, the authors intend a forward-looking focus that emerges from the fact that research and teaching with literary audio on this scale promises the possibility of paradigmPanel

Proposal for DH2020 shifting engagement with audio objects that have been, by and large, beyond the consideration of methodologies central to the discipline of literary studies and beyond the primary concerns of research librarians and archivists. An exciting dimension of SpokenWeb is the collaborative enterprise of digital humanists, cultural scholars, literary scholars, archivists, librarians, and sound technicians coming together to generate questions and solve problems beyond the capacity of any one of those fields of expertise. Such an undertaking invites the revisioning of processes of technical development, preservation and access strategies, metadata and systems coordination, rights management, pedagogical methods, student research and training, events curation and community outreach, and overarching concerns of project management and governance. With such revisioning comes the opportunity and responsibility to consider, or reconsider, the ethical dimensions of each set of project activities.

Our proposal to explore ethical case study scenarios in relation to SpokenWeb research activities, approached as a multi-authored work with representatives from different constituencies of our research partnership, stems from the governance structure of the network. The governance structure is designed to provide representation of the key research areas of the SpokenWeb network, namely, literary methods, data and metadata, computational analysis, pedagogy and training, community outreach and events, and reserves additional positions for student representation. We will make use of this structure to organize the papers comprising our presentation, and will conclude with some framing observations about the important function of project management and governance in relation to the ethical scenarios discussed.

Paper 1

"Ethical Considerations of Rights, Access and Metadata Management of Digital Audio Assets"

Annie Murray (University of Calgary) and Jason Camlot (Concordia University)

Determining the rights holder for an archival audio recording can present challenges to archivists and researchers alike. In many cases, collections of audio recordings may have scant labelling on them, and in many archival collections, descriptive information may be at a basic level. Normally, physical examination of the cassette can yield some information about the possible rights holder; however, listening to the digitized recordings can bring about a fuller awareness of the specific copyright information needed to pursue permissions. The individual or organization who made a given recording is in most cases the rights holder. That is to say that the speaker or reader on the recording is in many cases not the rights holder. To what extent might the speaker or performer be informed of or involved in the permissions process for a recording? How would an institutional archive approach rights management compared to an artist-centred community? Rights management must transcend legal documents and procedures, and take into account relationships amongst archival, library, researcher, and literary stakeholders. Since our project intends eventually to provide open public access to this entire body of recordings, rights management activities need to take into account affective and ethical considerations that arise in order to ethically share archival recordings. Rights and permissions management of time-based media assets that document historical performances and speech acts of individuals before and within communities of the past demands engagement with the implications of replaying such events within communities of the present.

In addition to such considerations surrounding the determination of permissions and access gradations for documentary audio recordings, we have been interested in reflecting upon the ethical implications of the nature and use of the descriptive metadata that we have produced and will continue to produce for the audio assets of the collections we are developing. During the first year of the SpokenWeb research program, a Metadata Task Force worked to develop a unique (yet linked) metadata schema for the description of documentary literary and humanities-oriented sound recordings held in both institutional and community-based collections. The schema is used by all team members to describe collections across the country. Interesting ethical questions have arisen both in relation to the establishment of specific metadata categories, and for the larger question concerning the degree to which the metadata we are producing should be made publicly accessible. Fields such as Genre, Content, and Statement of Responsibility have raised interesting questions about the impact of disciplinary presuppositions (in the case of genre), transcription (in the case of content) and definitions of "authorship" and communities of production (in the case of responsibility). We will identify and propose parameters of discussion for these three examples of metadata production and the ethical issues they raise. Finally, we will describe some pointed ethical scenarios pertaining to the use of metadata once it has been collected and made ready to share. While our first position has been (and remains, in principle) to make all SpokenWeb metadata open, we feel it is worth exploring possibilities and scenarios in which the descriptive metadata we are producing about documentary sound recordings

may itself raise questions of ethical disclosure. While the argument that the open sharing of the descriptive metadata we produce should, legally, be allowable because such information consists of descriptive and reported facts about artifacts that are linked to, yet separate in substance from, the artifacts that may be subject to more complicated rights requirements, ethical questions remain concerning whether the parties that are captured in documentary audio collections wish their past activities, associations and locations to be known and be made publicly accessible through openly shared data that is designed to be linked and discoverable. For example, description of events that make intensive use of VIAF and Wikidata to establish association links between scenes and individuals, without details concerning the nature of past associations, may raise issues about what Erving Goffman once referred to as "Stigma management" in the present. This discussion will be pursued with reference to work on the ethics of metadata by Brody 2003, King 2009, Sinha 2010, Farnel 2018, Laiola 2018, McMahan and Denaxas 2019, and others.

Farnel, Sharon (2018) "Metadata as Data: Exploring Ethical Metadata Sharing and Access for

Indigenous Resources Through OCAP Principles," Proceedings of the Annual Conference of CAIS /

Actes du congrès annuel de l'ACSI, ISSN 2562-7589, DOI: https://doi.org/10.29173/cais974

Goffman, Erving. Stigma: Notes on the Management of Spoiled Identity. New York: Simon and Schuster, 1963.

McMahon, Christiana & Spiros Denaxas (2019), "A novel metadata management model to capture

consent or record linkage in longitudinal research studies," Informatics for Health and Social Care

44:2, 176-188, DOI: 10.1080/17538157.2017.1364251

Sinha, Hans P. "The Ethics of Metadata: A Critical Analysis and a Practical Solution," The Maine Law Review 63.1 (2010): 176-262.

Whitcomb Laiola, Sarah. "Markup as Behavior toward Risk: Reforming the Metanarratives of

Metadata through Susan Howe's Metafictive Poetics." American Quarterly, vol. 70 no. 3, 2018, 561-

587.Project MUSE, doi:10.1353/aq.2018.0035.

Paper 2

"Closer Listening: or, beyond the 'anonymity of a murmur'"

Deanna Fong (Concordia University), Michael O'Driscoll (University of Alberta)

"We can easily imagine a culture where discourse would circulate without any need for an author. Discourses, whatever their status, form or value, and regardless of our manner of handling them, would unfold in the anonymity of a murmur."

Foucault, "What is an Author?"

When Michel Foucault conceived of an archaeology of knowledge, predicated on the emergence of "statements" that, as Giorgio Agamben famously challenges in the archival ethics of his Remnants of Auschwitz, are events void of subjectivity, or when Foucault simultaneously contended for the death of the author as a mere subject-function unnecessary to the circulation of literature, he could hardly have imagined a vast digital archive of poetic discourse comprised entirely of the recorded voices of successive generations of creative writers. Amongst its many multi-disciplinary implications, the SpokenWeb Project promises, in moving from page to stage, a paradigm-shifting invitation to consider the digital audiotext as a fundamental medium of literary-historical analysis. In turning our attention from the printed, written text to the voiced performance, the SpokenWeb archive not only prompts new methodological questions for literary analysis, it also insists upon a revived relationship to the human being occupying the reanimated subject-position of "author." Indeed, the audio objects at the centre of this collective enterprise-with their haunting disjointedness of the voice's spectral presence-pose pressing questions of relation-of ethical relation, obligation, inheritance, and responsibility-between speaker and listener.

This paper moves toward, or contends for, an ethics that is predicated on listening as intentional and affirmative of the speaker's embodied subjectivity-how we might listen to and for the grain of voice, the breath, the murmur, the semiotic rather than symbolic articulations of body-the openness of which runs counter to the hermeneutics of suspicion that governs conventional practices of textual analysis. As Mladen Dolar contends in A Voice and Nothing More: "The voice can be located at the juncture of the subject and the Other, just as it was before, in a different register, placed at the intersection of body and language, circumscribing a lack in both." The speaker's embodied subjectivity, foregrounded in the audio artifact, also points us to a set of material questions that have to do with the medium's "counter-affordances" as a register of public speech. What can and can't be said on the record, and by whom? Here, we argue that while audio recordings open us up to the stochastic noise of live speech and performance that these are also disciplined by the act of recording, and that this discipline unevenly applied to different subjects based on their relation to dominant power. In this manner, an ethics of archiving digital audio must also consider the material gaps and silences that exist within individual artifacts (in the form of cuts and

edits, self-censorship, and muted or inaudible voices) as much as within larger collections (representation, institutional prioritization, preservation practices, etc.).

Agamben, Georgio. Remnants of Auschwitz: The Witness and the Archive. New York: Zone Books,
1999.

Dolar, Mladen. A Voice and Nothing More. Cambridge, Ma.: MIT Press, 2006.

Erlmann, Veit, ed.. Hearing Cultures. Essays on Sound, Listening and Modernity. Oxford: Berg, 2004.

Fiumara, Gemma Corradi. The Other Side of Language: A Philosophy of Listening. London:
Routledge, 1990.

Foucault, Michel. The Archaeology of Knowledge. Trans. A.M. Sheridan Smith. New York: Routledge, 1972.

Michel Foucault, 'What is an Author?', in Paul Rabinow, ed., The Foucault Reader. New York:
Pantheon, 1984. 101-120.

Paper 3

"Automated Annotation, Metadata Production, and Scenarios of Difficult Content"

Tanya Clement (University of Texas at Austin), Liz Fischer (University of Texas at Austin)

Audio Visual collections, even when digitized, are difficult to search and analyze. Often, AV comes with limited metadata. Consequently, information professionals have to listen to items in real time and add information for indexing. Since enhancing searching, visualizing, and analyzing audio collections is often reliant on some level of metadata, an item that has not been vetted will often remain "hidden". Libraries and archives, after all, are not want to make cultural heritage items that might potentially have privacy or security issues openly available online. In such cases, libraries have begun to consider using machine learning and other means of automatically generating metadata. Automatically generating metadata introduces ethical concerns, however. Jen Guiliano and Carolyn

Heitman use the term "difficult heritage" to explore "the aggregation of cultural records related to colonialism and its impacts on Native peoples" in their desire to "reconcile concerns of documenting the past with the need to confront the genocidal practices from which most of these records result". In the context of SpokenWeb, such concerns are also relevant, though differently oriented. Many of the recordings include well-known poets such as Allen Ginsberg, Robert Creeley, and Muriel Rukeyser, but these poets also appear in recordings of panels or in the context of classroom encounters where unknown or unnamed voices also linger. Further, speakers may use language that is flippant or unconcerned at best or offensive at worst about topics such as race or gender in ways that are not reflective of more current perspectives and conversations. These incongruities are not new; they have always been part of our difficult heritage, but this discussion will articulate which ethical concerns have come to the fore when we use machines to ingest, sort, analyze, and represent sensitive materials in audio collections.

Jennifer Guiliano and Carolyn Heitman, "Difficult Heritage and the Complexities of Indigenous Data,"

Journal of Cultural Analytics. August 13, 2019. doi: 10.22148/16.044

Geneva Henry. "Research Librarians as Guides and Navigators for AI Policies at Universities." Research Library Issues, no. 299 (2019): 47-65. https://doi.org/10.29242/rli.299.4.

SAA Core values statement and Code of Ethics. Society of American Archivists. https://www2.archivists.org/statements/saa-core-values-statement-and-code-of-ethics

Paper 4

"Pedagogy, Mentorship, Labour in Digital Humanities Research Partnerships"

Karis Shearer (University of British Columbia, Okanagan), Emily Murphy (University of British Columbia, Okanagan).

A substantial proportion (~67%) of SpokenWeb SSHRC Partnership funding is dedicated to graduate and undergraduate students, whose training and research contributions to the SpokenWeb partnership form a core part of our mandate. The SpokenWeb Pedagogy Task Force has viewed this as an opportunity to develop mentorship and supervision best practices, building upon those outlined in "The Student Collaborators Bill of Rights" and with attention to findings and recommendations of "Student Labour and Training in the Digital Humanities" (Anderson et al). The integration of students into a large-scale digital humanities research network raises some key questions regarding mentorship, training, and graduate vs undergraduate student labour. For instance: what counts as "work" on SpokenWeb? How do we describe work the students are doing? Is it possible to ensure practices around student labour are addressed with consistency across the SpokenWeb network of Canadian and US American institutions? What kinds of iterative conversations are necessary to ensure work is meaningful to students? Murphy and Shearer consider the organization of roles and relations across the research network, with particular attention to the role of undergraduate trainees. What kinds of ethical issues emerge when involving undergraduate trainees in research roles?

Anderson, Katrina and Lindsey Bannister, Janey Dodd, Deanna Fong, Michelle Levy, Lindsey Seatter.

"Student Labour and Training in the Digital Humanities." Digital Humanities Quarterly 10.1
http://www.digitalhumanities.org/dhq/vol/10/1/000233/000233.html
Stone, Sandy. "Refuse closure; Insist on situation; Seek multiplicity." http://www.deeplab.net/aboutus
The Student Collaborators Bill of Rights (UCLA). https://humtech.ucla.edu/news/a studentcollaboratorsbill-of-rights

Paper 5

"The Ethics of Literary Curation, Archival and Actual"

Jason Camlot (Concordia University), Klara du Plessis (Concordia University)

Literary event curators adopt a range of curatorial approaches, varying from extremes such as Sarah Longair's notion of "curatorial authority" ("Cultures of Curating")-an intimate knowledge of work presented and a predefined context for its public representation-to Hans Ulrich Obrist's organic model of providing a space in which experience is generated according to the individuals presenting or interacting with works (Ways of Curating). Positioned as mediator between authors and audience, however, the literary curator accepts, by default, a radical responsibility towards both the authors they represent and towards the audience they expose to literary performance. Both as historical or archival study and as continued praxis for contemporary literary events, poetry reading curation hinges on an ethical discussion of relational dynamics as manifested in literary dissemination. This paper takes two pairs of scenarios--each comprised of one archival and one recent experiential example--to frame the ethical concerns surrounding the curation of literary events.

As a first case study, du Plessis follows the digital audio archives of Véhicule Art Inc--a bilingual, interdisciplinary artists' run centre and exhibition space that hosted a weekly poetry reading series and quarterly open mic marathon events in Montreal, 1972-1983--as well as experimental, contemporary Deep Curation readings--an approach which places poets' work in deliberate dialogue with each other and heightens the curator's agency toward the poetic product. As a curatorial strategy, the former prioritizes a self-conscious preference for literary novelty and unrestrained openness, offering all agency to the authors as to what they present to the audience. Arguably, this (along with a wide range of cultural and sociological factors inherent to the 1970s in North America) frequently results in an unfiltered performance of sexist, misogynist, rhetoric; this work often signals sexually violent content, especially, when mapped onto a contemporary listening audience. As a corrective, a Deep Curational approach scripts exactly what the authors will present, creating a thematic and theoretical framing for the reading, while also retaining curatorial agency for the curator. Although authors consent to this process, dynamics of agency between curator and authors undergo a radical shift as the curator now enacts an authorial role.

As a second case study, Camlot presents brief examples from SpokenWeb's first "Performing the Archive" series, developed by staging public readings in which poets who read in Montreal as part of the Sir George Williams Poetry Series (1965-1974) return to read alongside their past archival selves. The illustrative examples in this case will include events involving Daphne Marlatt, Diane Wakoski, George Bowering and David McFadden. Mixing live reading with readings from the archive results in an affective performance of the relationship between the modes and meanings of performance in a past cultural and sociological context and the present one. Further as the live reader stands silently on stage while the archival clips are played, it also renders the act of listening a mode of visible performance for the audience. Such performances function as a lo-fi versions of what Steven Benford and Gabriella Giannachi have described as acts of "performing mixed reality." Staging listening to the archive as a form of live performance can also be understood to open a temporal portal (to use a term from Wolfgang Ernst) between two ethico-cultural moments, and incites reflection upon the meaning of literary performance in different historical contexts. Each "Performing the Archive" event is followed by a period of discussion that explores the meaning of the temporal encounter that such a curated performance involving archival audio, entails, and, inevitably, raises interesting questions about the ethics of activating archives in distinctive, performative ways, and the different aesthetic and ethical meaning of literary performance in changing historical contexts.

Beneford, Steve, and Gabriella Giannachi. Performing Mixed Reality. Cambridge, MA: The MIT Press, 2011.
Ernst, Wolfgang. Sonic Time Machines: Explicit Sound, Sirenic Voices, and Implicit Sonicity. Amsterdam: Amsterdam University Press, 2016.
Longair, Sarah. "Cultures of Curating: The Limits of Authority." Museum History Journal (2015) 8:1, 1-7.
Obrist, Hans Ulrich and Asad Raza. Ways of Curating. New York: Faber and Faber, 2014.

Paper 6

"Project Management, Governance as Causes and Mitigators"

Yuliya Kondratenko (Concordia University) and Jason Camlot (Concordia University)

This paper presents scenarios concerning ethical considerations surrounding the management of

large scale, interdisciplinary DH projects, and how governance structures and processes may help mitigate conflicts across a research network.

The Project Management Institute's (PMI) Code of Ethics and Professional Conduct states that when project management practitioners are attempting to uphold principles of fairness, they should aspire to "...constantly reexamine [their] impartiality and objectivity, taking corrective action as appropriate." One of four foundational values of PMI Code (the other three are responsibility, respect, and honesty), this principle translates into the need for a consistent practice that aims to "...recognize when we (project management practitioners) have conflicted loyalties, and to identify when we are inadvertently placing ourselves or others in situations of conflict-of-interest."

How might this principle of fairness be enacted and applied in the context of a large and diverse research network with a wide range of stakeholders, a governance structure that is composed of humanities researchers, librarians, and community partners, all with competing interests. Take librarians, for example. On the one hand, they have a duty of loyalty to their own institution when performing their functional role. On the other hand, when serving in their role as a member of a research project (as a member of the Governing Board or a project-oriented Task Force), they may be asked to reconsider, reexamine or reshape policies and procedures that they have been hired to uphold in their institutional role. The project manager in such a scenario ends up enabling and perhaps even inciting a conflict-of-interest situation, which represents a conflict-of-interest in itself, for the project manager, as it undermines a key ethical principle of the PMI's Code of Ethics and Professional Conduct.

Figure 2. SpokenWeb Network and Governance Diagram

The establishment of a well designed governance structure can help to mitigate such scenarios of conflict of interest (and other forms of conflict). By identifying core features of the constituencies involved in the research program, ensuring representation of those constituencies on the Governing Board, and Task Forces, and especially by collaboratively establishing guideline and process documents that may be referred to for decision making in areas that represent regular points of tension between the goals of a network project, and the goals of a participating institutional or community partner, scenarios of conflict can be managed in such a way that they do not undermine the advancement of project aims, and to the satisfaction of the interests of the partner. Importantly, creating scenarios for collective reflection upon the complexity of our loyalties and relationships is a productive approach. Staging discussions about agreed upon methods of collaboration -- for example, the use of cross institutional Task Forces designed to advance specific goals -- is a useful and neutral way to reflect upon the effects of institutional loyalties in relation to shared goals of the research program. This paper presents scenarios of such conflict and resolution for consideration and discussion.

A Guide to the Project Management Body of Knowledge (PMBOK Guide). 6th ed. Newtown Square, Pa.: Project Management Institute, Inc, 2017. Print.

PMI's Code of Ethics and Professional Conduct: https ://www.pmi.org/-/media/pmi/ documents/public/pdf/ ethics/pmi-code-of-ethics.pdf? v=5b0f0983-6467-4d7d -9aae-577fbac4d4b3&sc_lang_temp=en

Janice Yu Chen Kung and Thane Chambers, "Implementation of a fee-based service model to university-affiliated researchers at the University of Alberta", Journal of the Medical Library Association 107.2 (April 2019). doi: http://dx.doi.org/10.5195/ jmla.2019.497

Towards an Indian Decolonial Digital Humanities

Dibyadyuti Roy
dibyadyutir@iimidr.ac.in
INDIAN INSTITUTE OF MANAGEMENT INDORE, INDIA

Nirmala Menon
nmenon@iiti.ac.in

INDIAN INSTITUTE OF TECHNOLOGY, INDORE, INDIA

Maya Dodd
maya@flame.edu.in
FLAME UNIVERSITY, INDIA

Arjun Ghosh
arjunghosh@hss.iitd.ac.in
INDIAN INSTITUTE OF TECHNOLOGY, DELHI, INDIA

Mayurakshi Chaudhuri
mchaudhuri@iitj.ac.in
INDIAN INSTITUTE OF TECHNOLOGY, JODHPUR, INDIA

Chiranjoy Chattopadhyay
chiranjoy@iitj.ac.in
INDIAN INSTITUTE OF TECHNOLOGY, JODHPUR, INDIA

Keywords: postcolonial, archives, community, history, south asia
Topics: Asia, Global, English, 20th Century, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), public humanities collaborations and methods, South Asian studies

Panelists include organizers and participants from the Digital Humanities Alliance of India (DHAI) Conference held in June 2018-the first digital humanities conference in India-as well as academics who have pioneered DH research and teaching projects in India. We examine why even after several years of academic inquiry DH leaves a faint footprint on humanistic and public-facing scholarship in India. The four papers in this panel will look at the intersections of philosophy, theory, pedagogy and practice: to conceptualize a postcolonial DH in India using examples of particular projects as case studies. Together the papers in this panel explore the possibilities of Digital Humanities in India along with its concomitant challenges-including technological, structural, institutional disparities as well as issues of access and mobility-and reimagines the potential for developing a decolonized praxis for Digital Humanities in the Global South.

No "Making", Not Now: A Distant Reading of Digital Humanities in India

Dr. Dibyadyuti Roy (Indian Institute of Management, Indore) and Dr. Nirmala Menon (Indian Institute of Technology, Indore)

The emergence of universities and postsecondary education in India from a colonial system that was "set up primarily for conducting examinations and awarding degrees, and not for undertaking research or even teaching" (Beteille para 2) has resulted in a system some have termed an educational apartheid: developing and sustaining the "ominous social pressure: [of]the idea that good students study science. Studying humanities is frowned upon as intellectually demeaning" (Gangopadhyay para 4). This ensures majority of public funding for research is pumped into the STEM fields while the Humanities have to make do with the crumbs. This in turn means that Humanities departments in universities often cannot even access disciplinary innovations in DH theories, methodologies and tools, let alone advanced training for faculty members. Therefore, the contextual relevance of the "making" versus "talking" debate raging across Anglo-American DH sites withers away in postcolonial spaces like India where DH, at least not yet, is not a term of "tactical convenience' (Kirschenbaum)

In acknowledging this ontological premise, our intervention tries to locate the inaugural *Digital Humanities Alliance of India (DHAI)* Conference in June 2018 within an epistemic genealogy of South Asian and more specifically Indian Digital Humanities interventions, which may act as a primer for envisioning the future of postcolonial and indeed Global Digital Humanities. We consider some of the reasons for the resistance to institutionalizing digital humanities within the larger university systems in India and conclude that the causes are historical but also perhaps the embedded rigidity and regimentation of university-curricula across disciplines. Digital Humanities, as defined by its purported inter- or transdisciplinary character, becomes a tough sell within Indian university ecosystems, often constrained by their colonial origins and the limitations of a postcolonial social imaginary. By highlighting key issues regarding digital affordances in postcolonial contexts and exemplifying digital humanities projects thatwere presented at the DHAI 2018 conference, along with representative digital interventionsbeyond the conference's ambit, we conduct what we may term a 'distant reading' of the digital humanities in India. In conclusion, we suggest the need for a postcolonial digital humanities that is deeply invested in cross-disciplinary exchanges and pedagogical innovation: one that interrogates hegemonic assumptions about both "Postcolonialism" and "Digital Humanities."

Do Postcolonial Intersections Matter?

Critical Epistemological Approaches to Digital Humanities and Mobility Studies in India

Dr. Mayurakshi Chaudhuri and Dr. Chiranjoy Chattopadhyay (Indian Institute of Technology, Jodhpur)

"Postcolonial theory" evidently is not a singular theoretical framework, but rather an interrelated set of critical and counterintuitive perspectives, including a complex network of interconnected concepts and heterogeneous practices that have been developed out of traditions of resistance to global historical trajectories of imperialism and colonialism (McLeod 2007). Furthermore, such counterintuitive perspectives and inquiry have undergone immense transformation by the digital revolution of the twenty-first century which has redefined relationships with and within the society, as well as sociological approaches to its analysis and interpretation in response to historical transformations over the past decades. A fundamental question remains as a natural corollary: How should this inquiry change with newer visibilities in the future, particularly for "digital humanists"? To examine these new challenges, we turn to an unavoidable global fact, omnipresent in historical and in contemporary times: human mobility. While sociological theorizing about mobility during the last quarter of the twentieth century was dominated by questions of relationship between mobility and neo-Western ideas of modernity (Sheller and Urry 2016), there has been a shift in discourse since the last decades with the advent of the digital revolution where postcolonial discourses have been re-crafting the very definition and production of mobility (for example, tourist sites).

However, while research on mobility studies over the decades have included objects of development whose actions are structurally determined, to cultural subjects across the continuum, one consistent truth has been that their experiences of structure, agency, and identity negotiations are inflected by gender, race, ethnicity, class, religion, sexuality and nationality, to name a few of the axes of differentiations. With this in background, this article seeks to bring from the periphery to the core the multiple connections between mobility, intersectionality, the colonial past and postcolonial dynamics particularly impacting Public Digital Humanities in India. Not surprisingly, prolonged realities of colonialism have produced historical and cultural consequences that remain integral to hegemonic mobility and contemporary lifestyles in India,and these axes of differentiations have had very long, contested historical roots (Loomba 1998).

For example, European colonialism in South-, Southeast- or East Asia, required, and made possible, inequalities of power that pivoted around apparently real yet ultimately imagined gender differences between the colonizers and the colonized, that gets manifested across societal and digital spaces. Therefore, examining mobility within the framework of postcolonial histories in India thus brings us to an important hinge between distant times and places - the linking is historical, cultural, material, and digital (Chaudhuri and Thimm 2018). Everyday negotiations of such mobility explicitly point to these pressure-points and movements across multiple geographic and social scales where the axes of differentiations are constantly negotiated and redefined. This opens up a plethora of questions that need immediate attention: (1) How are such postcolonial mobility scales (local, regional, national, digital) re-constructed? (2) How do these postcolonial scales inflect and are inflected by the various axes of differentiations; and most importantly, (3) How have these negotiations inflected postcolonial realities of mobility and lifestyles in digital India? That is, how are the new visibilities of postcolonial realities crafted?

This paper includes broader debates of these new visibilities of postcolonial experiences and lifestyles and en route, reflect on the importance of theorizing social position, power relations, individual's identities, and their accompanied complexities in understanding emerging postcolonial patterns that inform, influence, and even inflect epistemologies of Public Digital Humanities in India.

Visualizing the Cultural History of South Asia

Dr. Arjun Ghosh (Indian Institute of Technology, Delhi)

The history of the various literary and artistic forms have hitherto been written by the close study of the works and activities of individuals or artistic movements. Even though such methods - qualitative in character - have produced authoritative histories, they have relied on the human ability to read and correlate texts. The computational turn makes it possible, through large datasets, to support as well as put to test readings by the "human eye". Schich et al used crowdsourced data of birth and death dates and locations to study patterns of migration of notable people across history in Europe and North America (Schich et al. 2014). Schich et al. sourced their data from Freebase, a crowdsourced database of notable people and things, the Getty Union of Artist Names and other sources.

In this project we visualize the data for South Asia to look at patterns of migration. In order to update the data used by Schich et al. we supplement it with data from Wikipedia biographies. The most influential persons from the Wikipedia data is extracted using a filter for pagevisit statistics of the particular Wikipage. From

Wikipedia we can build a bigger dataset that includes birth and death data for people from different sectors of society. We map this data to reveal the different locations which acquire importance in different sectors across the 19th and 20th centuries. While migration studies relying on census and other macro level data concentrates on studying the migration patterns for labour and skilled workforce in various industrial sectors(Dupont 2000, Abhishek et al. 2017), working through crowdsourced data allows us to identify the changing patterns of migration among artists and creative persons, administrators, sports people, academicians and scientists and other elite sections of the population. Our analysis reveals that the erstwhile colonial capital Calcutta started losing importance as a cultural hub in the 1950s and 60s when a large number of artists from the western Punjab settled in Bombay. In recent years the southern cities of Chennai and Bangalore has been growing in attractiveness for artists in Southern India. Pune has also grown in attraction for artists and academicians in comparison to Mumbai. While different urban centres have changed positions as hubs for different sectors, Delhi has grown as an attractive settlement across sectors since the 1990s.

Creating Digital Archives in India: The Public's Record

Dr.Maya Dodd (FLAME University)

In an attempt to challenge the inscrutability of the state, the Right to Information (RTI) Act of 2005 has been critical in creating a public record and making government information visible to India's citizens. It is no coincidence that such a right was not only a function of democratic maturity but also a function of technological possibility. Without oversimplifying the point about transparency and accountability, the ease that digital access has granted in making an RTI request assumes the existence and possibility of digital archives for civic convenience. Since such archives are often being created at the time of the ask, the creation of digital registers of information is essentially a collaborative effort.

Sites like governmentdata.gov and citizen led efforts like open data meets are both enabling thecreation of parallel registers and repositories that often contest or carry forward the digital data sets of governance. Similarly, in the case of public history, in addition to officially sanctioned preservation projects, citizen led and crowdsourced efforts to populate public records are enabled by digital media. Thus, we see portals like the People's Archive of Rural India (PARI) and theIndian Memory Project that are both a function of the digital medium as well as a product of a new democratic impulse to arrange information publicly. Understanding how these crowdsourced contributions multiply public records and further the cause of digital access can lead to insights in participatory governance and enhancing the public's access to the workings of the Indian state. When placed alongside other efforts to create digital archives (institutional archives, digital news and online museums for instance) the shift this indicates for public culture bears great promise for multiplying points of view, so necessary to a successful democracy.

Designing and Gaming: Doing Digital Humanities in India

Padmini Ray Murray

p.raymurray@gmail.com
DESIGN BEKU COLLECTIVE

Aditya Deshbandhu

adityad@iimidr.ac.in
INDIAN INSTITUTE OF MANAGEMENT INDORE, INDIA

Siddhartha Chakraborti

siddharthachakraborti@gmail.com
ALIGARH MUSLIM UNIVERSITY, INDIA

Keywords: postcolonial, gaming, south asia
Topics: Asia, Global, English, 20th Century, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), public humanities collaborations and methods, Games studies, South Asian studies

The panel will discuss a 'meta-understanding' of gaming culture and design ethos within postcolonial contexts and spaces. This panel will discuss specific DH projects especially in the areas of Gaming and Design to offer varied perspectives and possibilities as well as challenge how such (existing) epistemes and practices are conceived and mediated by the Global North. The three papers in this panel together situate these emerging DH practices and projects as complex rhizomatic activities whose theorizations and logic are yet to be narrated or articulated with any cohesiveness, primarily due to tactical reasons: which allow for a dispersed examination of gaming and design cultures in the heterogenous (neoliberal) settings of India. Secondly, they also offer us insight into how these inherently technological activities are re-shaped and

remediated by the users thus forcing academics and researchers to develop new ways of understanding of these artifacts.

Forums

Exploring the Undiscovered Contours of DH

Amanda Marie Licastro
amanda.licastro@gmail.com
Stevenson University, United States of America

Lee Elaine Skallerup Bessette
ls1335@georgetown.edu
Georgetown University

Zachary N Whalen
zwhalen@umw.edu
University of Mary Washington

Anne B. McGrail
mcgraila@lanecc.edu
Lane Community College, Eugene, Oregon US

Keywords: pedagogy, marginalization, digital studies, infrastructure, curriculum
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, digital activism and advocacy, Education/ pedagogy

How can scholars on the margins of DH articulate their work in DH publications, grants, and other professional and disciplinary outlets? In this interactive forum, we aim to explore inclusion--or exclusion--in what counts as "digital humanities" among scholars across disciplines, institutional contexts, and employment statuses. We begin by surveying audience members about the alternative ways that they represent their digital work and their different institutional contexts. We then ask participants to explore how esoteric terms such as "digital humanities" may be illegible to administrators and the public and the effects of this illegibility on their pedagogy and professional work. After collectively articulating the problem as it stands today, forum leaders will facilitate a problem-solving conversation that might begin addressing the issue.

Who needs tool directories? A forum on sustaining discovery portals large and small

Laure Barbot
laure.barbot@dariah.eu
DARIAH

Quinn Dombrowski
qad@stanford.edu
Stanford University, United States of America

Frank Fischer
frank.fischer@dariah.eu
Higher School of Economics, Moscow, Russia

Geoffrey Rockwell
grockwel@ualberta.ca
University of Alberta, Canada

Lisa Spiro
lspiro@rice.edu
Rice University, United States of America

Keywords: directories, discovery, infrastructure, tools, sustainability
Topics: Global, English, Contemporary, digital research infrastructures development and analysis, meta-criticism (reflections on digital humanities and humanities computing), Humanities computing

Digital humanists broadly agree that tool directories are a good and valuable thing, worth building and maintaining, but there is no sustainability model. This forum aims to move beyond platitudes and interrogate the value of directories and possible models for sustaining them. For whom are tool directories valuable, and in what context? Tool directories require ongoing attention in order to remain relevant -- and more technically sophisticated directories face infrastructure maintenance costs as long as the directory remains online and functional. Some directories have adopted a crowdsourcing model to address content updates, translation, and other necessary functions once grant funding runs out. As unpaid labor increasingly becomes

an area of attention and concern for digital humanities, one is left to ask, how ethical are directories, particularly when this volunteer labor is at particular risk of being lost through fragile infrastructure? At a certain point, time and funding are a zero-sum game: are directories actually worth it?

This forum is organized by individuals representing a range of DH directories, spanning from 2002 (Geoffrey Rockwell's TAPoR) to 2020 (Frank Fischer and Laure Barbot representing the Marketplace developed within the SSHOC project), along with the defunct DiRT (Lisa Spiro and Quinn Dombrowski). We will share in advance a brief white paper with case studies and provocative questions, and elicit discussion via Twitter and a mailing list. For the forum itself, we will briefly introduce the context for the discussion to accommodate participants who have not read the white paper (15 minutes). We will then organize the forum attendees into breakout discussion groups, each focused on 1-2 agreed-upon questions (15 minutes). These breakout discussions will last 30 minutes, and each group will have one participant report back to the larger group, which will lead into a general discussion (20 minutes). We will close the forum with a short synthesis and ask participants to suggest next steps (10 minutes).

The forum will help the organizers grapple with the difficult decisions that fall out from the "directory paradox", where the DH community's praise of directories is wildly incommensurate with the interest or resources available for sustaining them. Similarly, we hope it will help others running directories at various scales (including in forms such as Libguides or lists on GitHub pages) return to their projects with a clearer sense of what they're doing, for whom, why, and for how long.

Questions:

1. How should we review tools?
    a. Validity of tools for designers' stated purpose
    b. Usability in classroom
    c. Applicability to certain research questions
2. Who are directories useful for?
    a. And to do what?
    b. Are directories the most cost-efficient way to do these things?
3. Who should be responsible for creating and maintaining directories?
    a. What would a directory look like that you personally would be willing and able to contribute to? Does that even exist?
4. How can they be sustained when the grant runs out?
5. How can students use them?
6. Do we need standardized ontologies?
7. How can we involve volunteers?
8. What are the different models for tool directories?
9. Why shouldn't we just let Google do it?

Sharing is Daring: Global Copyright Laws, Knowledge Protection, and the Impact on Open Digital Humanities

Vanessa Hannesschläger
vanessa.hannesschlaeger@oeaw.ac.at
Austrian Academy of Sciences, Austria

Erik Ketzan
eketza01@mail.bbk.ac.uk
University of London, United Kingdom

Kim Nayyer
kpn32@cornell.edu
Cornell University, USA

Padmini Ray Murray
p.raymurray@gmail.com
Independent Researcher, India

Walter Scholger
walter.scholger@uni-graz.at
University of Graz, Austria

Benito Trollip
benito.trollip@nwu.ac.za
South African Centre for Digital Language Resources, South Africa

Keywords: Copyright, Indigenous Knowledge, Cultural Heritage, Open Digital Humanities
Topics: Comparative (2 or more geographical areas), Global, English, 20th Century, Contemporary, copyright, licensing, and permissions standards, systems, and processes, digital access, privacy, and ethics analysis, First nations and indigenous studies, Law and legal studies

It is now beyond question that opening up access to scholarly knowledge is a key value of the academy in the 21st century (cf. Berlin declaration). The Open Science paradigm however is not only restricted to facilitating democratic access to knowledge but also accommodates a set of key values of present-day knowledge creation such as fairness, transparency, equality and increased rigour and accountability in scholarly activities. Open Science therefore involves a cultural shift in the science ecosystem where research workflows are made transparent and subject to greater scrutiny from the early stages of protocols and planning through publication and data sharing. However, copyright laws and other legal frameworks often hinder or even prevent the Opening Up of research (be it research in progress, research data and tools, or research outputs and results).

In this forum, we would like to discuss the open sharing and making available of research and knowledge in general, and the global legal restrictions and privileges around copyright and licensing in the context of Open Digital Humanities and, specifically, DH cooperation with cultural heritage institutions on a global scale. Legal experts with in-depth knowledge of various copyright and indigenous knowledge protection systems across the globe will share their perspectives on the similarities and differences between legal systems that facilitate or hinder the Opening Up of humanities data. As these data are often provided (or protected) by cultural heritage institutions, their perspectives as stakeholders in this realm will be considered with special attention. Thus, not only copyright issues will be discussed, but also adjacent areas of the law that are relevant to the possible degree of Openness of humanities research, such as archival laws, personal data protection, and others. In addition, a special focus will be put on the protection of indigenous knowledge, on the question of how it is practiced and preserved by the communities, and the strategies that can be employed to support researchers on the one hand, and protect the interests of the affected communities on the other.

The forum is intended to be an open "fishbowl". The forum facilitators will each provide a short input of max. 3 minutes, briefly introducing themselves and naming the legal system they are experts in (Canada, Europe, India, South Africa and the USA, respectively) and one key area of expertise (e.g. indigenous knowledge protection, open licensing, cultural heritage data). Subsequently, the audience will be invited to propose topics for discussion and pose questions. All participants will be encouraged to contribute to the discussion, and the experts will answer concrete legal questions that come up. A moderator will ensure the balance of expert / participant contributions and the productivity of the discussion. The most prominent and pressing issues will be collected and summarized in a blog post jointly authored by the expert panel. It will be published on the blog of the DARIAH-EU working group "Ethics and Legality in Digital Arts and Humanities" (ELDAH).

References

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, https://openaccess.mpg.de/Berlin-Declaration

Ethics and Legality in Digital Arts and Humanities (ELDAH), https://eldah.hypotheses.org/

Quantitative Text Analysis: Intersections with Feminism, Postcolonialism, Critical Race Theory and Queer Theory

Katherine Bode
Katherine.Bode@anu.edu.au
Australian National University, Australia

Alison Hedley
afhedley@gmail.com
McGill University, Canada

Natalie M. Houston
Natalie_Houston@uml.edu
University of Massachusetts - Lowell, United States

Anouk Lang
anouk.lang@ed.ac.uk
University of Edinburgh, United Kingdom

Emily Murphy
emily.murphy@ubc.ca
University of British Columbia, Canada

Rianna Walcott
rianna.walcott@kcl.ac.uk
King's College London, United Kingdom

Keywords: quantitative text analysis, text mining, feminism, postcolonialism, critical race theory
Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), text mining and analysis, Feminist studies, Literary studies

This forum brings insights from bodies of theory attentive to the material effects of hierarchies and difference - including but not limited to feminism, critical race theory, queer theory, and postcolonialism - to bear on quantitative text analysis. It seeks to bring into view the power relations that inform processes of corpus construction and that underlie the computational models, analytical frameworks and platform architectures through which born-digital texts are obtained. Questions the forum will open up include how the positivistic claims of objectivity associated with quantitative text analysis might be resisted and subverted, how quantitative text analysis might interrogate and dismantle hierarchies rather than reinforcing them, and the kinds of infrastructural, institutional or systemic changes that are needed to instantiate alternative modes of text analysis.

Ensuring Software Quality and Sustainability of DH Projects

Julia Damerow
jdamerow@asu.edu
Arizona State University

Diego Siqueira
diego.siqueira@uni-muenster.de
Service Center for Digital Humanities, University of Münster

Robert Casties
casties@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Malte Vogl
mvogl@mpiwg-berlin.mpg.de
Max Planck Institute for the History of Science

Keywords: research software engineers, software quality, research quality, community building
Topics: Global, English, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), software development, systems, analysis and methods, Computer science, Humanities computing

Many Digital Humanities projects involve the development of software or scripts. However, the development of software is often not understood as proper academic output that counts towards academic performance. In addition, software developers in DH often lack formal software engineering training, although not all concepts from the software development domain can be fully applied to DH research projects. This often leads to unmaintained or poorly written software that is hard or impossible to reuse by others. We propose a forum centered around the question of how the DH community can ensure code quality and sustainability beyond the limited and often temporary funding of DH projects. The discussion would be focusing on questions such as how the community can provide proper credit for coding related work, procedures to improve code quality and code sharing, and how scholars entering the field can receive training and support in programming related topics.

Towards a Multilingual DH

Geoffrey Bacon
University of California, Berkeley

Svenja Gülden
sguelden@uni-mainz.de
AKU Project

Roseline Agunbiade
agunbiademodupe@gmail.com
Federal University of Technology, Akure

Lorella Viola
lorella.viola@uni.lu
Utrecht University

Keywords: language, multlingual, OCR
Topics: Global, English, BCE-4th Century, 15th-17th Century, Contemporary, digital ecologies and digital communities creation management and analysis,

optical character recognition and handwriting recognition, Humanities computing

The Anglophone-centricity of digital humanities continues to be an impediment to DH scholars who work on, or teach with, materials in other languages. Tools and tutorials are almost all designed to work with English, and while it may be possible to adapt them for use with other languages, it's not always obvious for a given language what adaptations are needed -- or even possible -- to bridge the language gap. This forum will be an opportunity for anyone with an interest in multilingual DH to take stock of the current state of the field as applied to any language(s) that they work with, discuss the impediments they face in doing multilingual DH, and identify potential collaborators for addressing these issues.

Using and Sharing Crowdsourced Data

Victoria Anne Van Hyning
vvanhyning@loc.gov
Library of Congress, United States of America

Samantha Nicole Blickhan
samantha@zooniverse.org
Zooniverse.org, Adler Planetarium

Heather Froehlich
hgf5@psu.edu
Penn State University

Keywords: crowdsourcing, public humanities, open data; data reuse
Topics: Global, English, 15th-17th Century, 19th Century, Contemporary, crowdsourcing, data publishing projects, systems, and methods, Humanities computing, Library & information science

Crowdsourcing projects are proliferating in the humanities and gallery, library, archive and museum (GLAM) world. These projects often produce complex datasets that provide new challenges and opportunities for practitioners across many disciplines. We need to have more conversations about how best to use and share this data, and to be open about the failures, as well as the successes of our crowdsourcing project designs and methodologies. This forum will speak directly to two of the conference themes this year: Digital Public Humanities and open data.

Sunny Days Sweepin' the Clouds Away: Critiquing the Digital Technologies Operationalized in Smart Cities

Constantine Gidaris
gidarisc@mcmaster.ca
McMaster University

Arun Jacob
arun.jacob@mail.utoronto.ca
University of Toronto

Angela Orasch
orascha@mcmaster.ca
McMaster University

Aaron Tucker
artucker@yorku.ca
York University

Keywords: smart cities, surveillance studies, facial recognition, political economy
Topics: English, North America, Contemporary, digital activism and advocacy, public humanities collaborations and methods, Cultural studies, Political science

This forum will highlight critical perspectives on theorizing the smart city by bringing together an interdisciplinary set of scholarship. Aaron Tucker examines the role of computer vision and facial recognition software in the smart city. Angela Orasch presents an analysis on the political economy of smart-city aesthetics vis-à-vis the processes of gentrification and displacement. Arun Jacob interrogates how digital mapping practices reveal the 'genealogies of power' operationalized in smart cities, and how activists can champion participatory democratic practices in these spaces. Constantine Gidaris looks at the negative impacts that systems led by racialized modes of digital surveillance and predictive policing and

technologies continue to have on marginalized people and communities.

The Open Lab: Digital Scholarship and the Question of Community

Matthew Nathan Hannah
hannah8@purdue.edu
Purdue University, United States of America

Brad Rittenhouse
bcrittenhouse@gatech.edu
Georgia Institute of Technology, United States of America

Jennifer Grayburn
grayburj@union.edu
Union College, United States of America

Sarah Connell
sa.connell@northeastern.edu
Northeastern University, United States of America

Brandon Walsh
bmw9t@virginia.edu
University of Virginia, United States of America

Keywords: digital scholarship labs, infrastructure, openness, public humanities, digital humanities
Topics: English, North America, Contemporary, curricular and pedagogical development and analysis, digital research infrastructures development and analysis, Humanities computing, Library & information science

Taking the conference themes of "openness" and public humanities, this forum will focus on the role of the digital scholarship lab within the larger university and broader community. Whereas we may assume such centers are hubs of activity and intellectual exchange, such spaces can also often be alienating to those not directly affiliated with the lab's range of operations. As leadership at five different labs within different departments and institutional contexts, we address strategies to foster and encourage openness, both within the university and toward the broader public. Despite our different universities and structures, we share a commitment to fostering healthy, sustainable, and welcoming labs, and we hope to engage the audience to think about and theorize openness as a central ethos in DH infrastructure.

Representation Matters: How Do We Promote a Radically Inclusive Digital Humanities?

Nathan Kelber
nkelber@gmail.com
JSTOR Labs

Hannah L. Jacobs
hannah.jacobs@duke.edu
Duke University Triangle Digital Humanities Network

Maggie Murphy
mmurphy@uncg.edu
UNC Greensboro Triangle Digital Humanities Network

Brooke Andrade
National Humanities Center

John Knox
UNC Wilmington Triangle Digital Humanities Network

Kemba N'Namdi
Triangle Digital Humanities Network Museum of Durham History

Melissa Lingle-Martin
Triangle Digital Humanities Network

Keywords: inclusion, representation, funding, networks, community-building
Topics: English, North America, Contemporary, meta-criticism (reflections on digital humanities and humanities computing), public humanities collaborations and methods, Education/ pedagogy

DH practitioners often stress a kind of radical inclusivity as foundational to the field. However, it

is clear that uptake, funding, support, and attention have been greatest in R1 institutions, leaving behind teaching institutions, liberal arts colleges, art and design schools, HBCUs and more. At the same time, DH has traditionally been very male and white. Digital humanities can be most fruitful, however, when it functions at intersections, connecting people of different backgrounds, institutions, and geographies. We bring together this forum, facilitated by members a growing DH network based in North Carolina, to give a space for DH practitioners to share methods and models for creating a radically inclusive Digital Humanities. We encourage participants doing DH outside of the R1 context to help us address important questions about inclusivity, scale, resources, institutional structures, representation, and community participation in new and expanding local and regional DH networks.

Asian American Digital Humanities: Building Radical Communities through Justice-Oriented Praxis

Anne Cong-Huyen
annech@umich.edu
University of Michigan, United States of America

Dhanashree Thorat
dt1349@msstate.edu
Mississippi State University, United States of America

Setsuko Yokoyama
setsuko@umd.edu
University of Maryland, United States of America

Arun Jacob
arun.jacob@mail.utoronto.ca
University of Toronto, Canada

Amardeep Singh
amsp@lehigh.edu
Lehigh University, United States of America

Keywords: Asian American Studies, Race, Transnationalism, Labor, Collaboration
Topics: AsiaGlobalEnglishNorth America20th Century, Contemporary, mixed-media analysis, public humanities collaborations and methods, Cultural studies, Media studies

This forum brings together Asian Americanist faculty, librarians, and students to unpack and frame research, pedagogy, and praxis in both digital humanities and Asian American studies: What does Asian/Am DH look like? What form might a community of practice in Asian/Am DH take? How best can we support Asian Americanists, whose field is historically grounded in community organizing and activism, in applying DH methods to their scholarship? As we address these questions, we will foreground how Asian/Am DH, like other ethnic studies informed DH praxis, centers concerns of race, social justice, transnationalism, and community. In the ethos of public digital humanities, one of the conference themes for DH2020, this forum emphasizes inclusion, care, community, and anti-racist collaboration.

Building new global DH communities: Africa and beyond

Juan Steyn
juan.steyn@nwu.ac.za
SADiLaR, South Africa

Tunde Ope-Davies
Obopeibi@unilag.edu.ng
University of Lagos, Nigeria

Felix Ameka
f.k.ameka@hum.leidenuniv.nl
Leiden University, The Netherlands

Sara Petrollino
sara.petrollino@gmail.com
Leiden University, The Netherlands

Keywords: Africa, global, participation, access, networks
Topics: Africa, Comparative (2 or more geographical areas), Global, English, Contemporary, digital

access, privacy, and ethics analysis, digital research infrastructures development and analysis, Education/ pedagogy, First nations and indigenous studies

DH has been expanding almost exponentially over the past few years, with the DH2019 conference being the biggest conference to date with more than 1000 participants. Despite this large number, broad representation of a global DH community is still lagging behind, with North-America and Europe still representing more than 80% of the participants despite international growth elsewhere, especially in Africa.

This forum aims to build on this foundation of DH2019 as well as the work done by GO::DH and aims to further facilitate discussions around the theme of how we can promote global DH intersections, accessibility and participation not just for African DH scholars but any independent scholar.

The forum will have three elements:

1. Pre-conference engagement with the DH community
2. The forum session at DH2020.
3. Dissemination of results as part of an enhanced publication

Discussion points will be facilitated through multimodal asynchronous contributions before, during (streaming) and after the conference.

FAIR Data Principles for the Humanities and Social Sciences: Sharing Knowledge and Perspectives Between Europe and Canada

Kevin Kee
kkee@uottawa.ca
Ottawa University

Arnaud Gingold
CO-OPERAS/OpenEdition, France

Nicky Agate
HuMetricsHSS, United States

Emilie Paquin
Erudit, Canada

Mark Leggott
Research Data Canada

John Simpson
Compute Canada

Jeff Moon
Portage/CARL

Keywords: FAIR principles, research data, Humanities, policies
Topics: Europe, English, North America, Contemporary, data publishing projects, systems, and methods, metadata standards, systems, and methods, Humanities computing, Library & information science

When it comes to open science and to defining pathways to achieve it, the specificities of research practices in the Humanities and Social Sciences are a challenge. HSS research fields are fragmented into many disciplines. Usually grounded in regional, national and linguistic communities, HSS research data are multifaceted, encoded in various formats, languages, and metadata schemas.

Through the presentation of CO-OPERAS' objectives and work plan, this forum proposal, driven by Kevin Key, dean of the Faculty of Arts of the University of Ottawa, will bring together European and Canadian stakeholders to engage in a trans-Atlantic discussion around open science and FAIR principles in the Humanities and Social Sciences (HSS).

More specifically, by tackling the topic with a specific focus on public and institutional policies, we will ask ourselves, and the public, what would be the conditions for the Canadian research community to efficiently join the movement.

Exploring Conferences as Sites of Public Digital Humanities in Times of Crisis

Roopika Risam
rrisam@salemstate.edu
Salem State University, United States of America

Sylvia Fernández
sferna109@gmail.com
University of Kansas, United States of America

Alex Gil
colibri.alex@gmail.com
Columbia University Libraries, United States of America

Keywords: public digital humanities, mobilized humanities, experimental humanities, digital activism
Topics: Global, English, Contemporary, digital activism and advocacy, public humanities collaborations and methods, Cultural studies, Media studies

Building on the facilitators' mobilized humanities work, including hosting hackathons for *Torn Apart/Separados* (http://xpmethod.plaintext.in/torn-apart/volume/1/) outside of the DH2018 conference, this forum explores how to effectively create space at conferences for public digital humanities that responds to political, social, and environmental crises. Participants take part in two activities: 1) an exercise doing mobilized public digital humanities and 2) a discussion of successes and challenges of the experience to articulate effective practices for undertaking mobilized public digital humanities at conferences. The outcomes of the forum will be disseminated through the Nimble Tents Toolkit (https://nimbletents.github.io), a set of how-to guides for public digital humanities, with all participants who wish to be named receiving credit. All conference attendees can fully participate, regardless of their humanities or technological skillsets, and can do so with the technology they have available to them - laptops, smartphones, tablets - or by working with a partner.

Demystifying ADHO, the Alliance of Digital Humanities Organizations

Glen Worthey
glen.worthey@gmail.com
University of Illinois at Urbana-Champaign, United States of America

Elisabeth Burr
burr-adho@dubhe.uberspace.de
Universität Leipzig, Germany

Brian Croxall
brian.croxall@byu.edu
Brigham Young University, United States of America

Isabel Galina Russell
igalina@unam.mx
Universidad Nacional Autónoma de México

Matthew Gold
mattgold@gmail.com
City University of New York Graduate Center, United States of America

Leif Isaksen
leifuss@googlemail.com
Exeter University, United Kingdom

Diane K. Jakacki
dkj004@bucknell.edu
Bucknell University, United States of America

Micki Kaufman
micki.kaufman@gmail.com
City University of New York Graduate Center, United States of America

Ernesto Priani Saisó
epriani@gmail.com
Universidad Nacional Autónoma de México

Ashley Sanders Garcia
asandersgarcia@g.ucla.edu
University of California, Los Angeles, United States of America

Sara Sikes
sara.sikes@uconn.edu
University of Connecticut, United States of America

Juan Steyn
zjsteyn@gmail.com
North-West University, South Africa

Keywords: ADHO, DH organizations, infrastructure

Topics: Comparative (2 or more geographical areas), Global, English, Contemporary, project design, organization, management, Humanities computing

This forum aims to "demystify ADHO," bringing together people with "insider knowledge" of ADHO based on their participation on multiple committees and in leadership roles in both ADHO and a few of its Constituent Organizations. These facilitators will give a series of lightning talks covering ADHO's organizational structure, finances, committees, and activities, and they will then open the floor for questions and discussion. We welcome even the most difficult questions and topics and will strive to be honest and open about ADHO's challenges.

Should We Burn it Down? The State of Librarians in Digital Humanities Pedagogy, From Uriah Heeps to Unicorns

Caitlin Christian-Lamb
caitlin.christianlamb@gmail.com
University of Maryland, United States of America

Nickoal Eichmann-Kalwara
nickoal.eichmann@colorado.edu
University of Colorado Boulder, United States of America

Ian Goodale
ian.goodale@austin.utexas.edu
University of Texas at Austin, United States of America

Sarah Melton
sarahmelton@bc.edu
Boston College, Unites States of America

Keywords: labor, pedagogy, collaboration, libraries, academia
Topics: English, North America, 20th Century, Contemporary, curricular and pedagogical development and analysis, meta-criticism (reflections on digital humanities and humanities computing), Education/ pedagogy, Library & information science

This forum aims to open up critical discussion about the roles of librarians in digital humanities (DH) pedagogy, and more broadly, how organizational structures and practices of the university writ large have 1) shaped the roles available for librarians within DH and 2) deepened a cultural divide between information professionals and other academic disciplines.

What does it mean to teach digital humanities as a librarian? What is library pedagogy today? What is the role of DH pedagogy in library-based DH centers, which may be in the library but not of the library? How does the neoliberal-led assessment-turn impact library pedagogy, particularly library DH pedagogy? What are the relationships between librarians and the rest of academe? This forum aims to address these questions, while also serving as a provocation on the relationships between digital humanities and library and information science, and the role of librarians connecting those spheres through pedagogy.

Pedagogy of the Digitally Oppressed Practicing Anti-colonial DH Pedagogy and Research

Arun Jacob
arun.jacob@mail.utoronto.ca
University of Toronto, Canada

Ashley Caranto Morford
ashley.morford@mail.utoronto.ca
University of Toronto, Canada

Kush Patel
kushpatel@avani.edu.in
Avani Institute of Design, India

Keywords: anti-colonialism, pedagogy, research methods
Topics: Global, English, North America, Contemporary, digital activism and advocacy, public humanities collaborations and methods, Education/ pedagogy

This forum invites scholars, creative practitioners, activists, and community members to collectively discuss and develop strategies for refusing the

damaging colonialities too often perpetuated within digital humanities teaching, learning, and research practices. Some of the topics that we hope to touch on in this discussion include: 1) how colonial ideologies and extractive research methods are naturalized within hegemonic DH principles and practices, 2) what anti-colonial DH pedagogies and insurgent research practices we might incorporate into individual contexts of digital humanities knowledge-making, especially given the uneven distribution of and access to digital infrastructures along the campus-community as well as the Global North-Global South spectrum, and, 3) how to sustain spaces for healing and community building within the realities of these uneven and dispersed infrastructures.

dh2020.adho.org