DATA REUSE AMONG DIGITAL HUMANITIES SCHOLARS:

A QUALITATIVE STUDY OF PRACTICES, CHALLENGES AND OPPORTUNITIES

by

Lina Marie Harper

A Thesis Submitted in Partial Fulfilment of the

Requirements for the Degree of

MASTER OF INFORMATION

in the School of Information Studies

ABSTRACT

Scholarship is more and more data-driven, and as digital tools continue to evolve, sound data use practices among scholars are now essential for scientific discovery. Data reuse is central to an emerging cultural push towards a more open way of doing science. The study investigates the challenges and opportunities in reusing research data among digital humanities (DH) scholars. Its findings may serve as a case study for how disciplinary practices influence the ways in which scholars reuse data. The aim of the study is to enhance current thinking and provide insight for information and library science practitioners at the intersection of DH and data reuse.

Data were collected using interviews. An analysis of semi-structured interviews with 12 DH scholars working in universities, research centres and cultural and heritage organizations internationally was performed. The study found that lack of time and resources, inconsistent data practices, technical training gaps, labour intensity and difficulties in finding data were the most challenging. Participants revealed a number of enabling factors in data reuse as well, and chief among them were collaboration and autodidacticism as a feature of DH. The results indicate a gap between data reusers and data sharers – low rates of sharing reduce the amount of findable and accessible data available for reuse. Both data reusers and data sharers must begin to see themselves as embedded into the research data lifecycle within the research infrastructure. The recommendation includes policy, education, and structural changes to the research culture, including data literacy, the development of self-paced RDM

training material, improvements to data discovery systems, rewarding data sharing, and creation of curation and preservation networks to help with data stewardship. The primary audience for this thesis is DH practitioners, librarians, funders and data repository managers and designers but findings around researcher motivations for data reuse may be helpful for researchers in the humanities and social sciences.

ACKNOWLEDGMENTS

I wish to express my deep appreciation to the scholars in my study, who generously shared their time and reflections with me. I also extend deep thanks and appreciation to the people who supported me in the successful completion of my master's degree and thesis. First and foremost: thank you to my supervisor, Dr. Stefanie Haustein. I am grateful for your support, expertise, and generosity. *Du kennst dich in der Wissenschaft wirklich aus und es war mir eine Ehre, von dir zu lernen*. Thank you also to professors, colleagues, scholars, librarians who guided me throughout my educational journey: Dr. Mary Cavanagh, Hélène Carrier, Dr. Constance Crompton, Dr. Kathleen Gregory, Dr. Mario Malički, Dr. Anton Ninkov, Dr. Jada Watson, everyone at ÉSIS (Evan Sterling for looking at early drafts) and at the ScholComm Lab (Juan Pablo Alperin, Alice Fleerackers), colleagues in the DRI (Erin Clary and curators, Mark Leggott, Lee Wilson, Jen Pecoskie, Victoria Smith, Alex Thistlewood), Scholar's Portal, SSHRC and LAC. Thank you to kind, critically minded friends in the information science community, locally and beyond. Thank you to my Concordia community of friends and professors, namely Matthew Hays, Cynthia Martin, Laura Roberts, and Shawna O'Flaherty.

I am grateful for and acknowledge the financial support from the Meaningful Data Counts project, Stefanie Haustein, the Coalition Publica scholarship program, the May Court Club of Ottawa, Lise Courchesne and Susan Harper. Je souhaite remercier les membres de ma famille et de ma communauté pour son appui et son encouragement :

Susan Harper (and Loreen) et la famille Courchesne (surtout Lise, Émile et Berthe). This MA was co-earned by Buster Brown. I also wish to underscore the importance and influence of historically-first data collectors: Indigenous Peoples. First Nations, Inuit, and Métis Nations have been researchers of the land for millennia, and their communities have been harmed by colonial and unethical research practices for decades. In solidarity, I affirm the importance of their data sovereignty, and the access to and possession of data that concerns them, in perpetuity. I acknowledge and give thanks for the land on which I live and play, where the majority of this research was conducted: the unceded territory of the Algonquin Peoples near the Kitigan Zibi Anishinabeg First Nation.

TABLE OF CONTENTS

DATA REUSE AMONG DIGITAL HUMANITIES SCHOLARS: A QUALITATIVE STUDY OF PRACTICES, CHALLENGES AND OPPORTUNITIES I

ABSTRACT II

ACKNOWLEDGMENTS IV

TABLE OF CONTENTS VI

LIST OF TABLES IX

LIST OF FIGURES X

LIST OF APPENDICES XI

CHAPTER 1: INTRODUCTION 1
1.1. Disciplinarity and Data Reuse 3
1.2. Research Objective, Questions, Rationale and Goal 6
1.3. Approach 7

CHAPTER 2: LITERATURE REVIEW 9
2.1. Defining Data and Data Reuse 10

2.2. Research Infrastructure and Data Scholarship 13

2.3. Overview of Data (Re)Use 16
2.3.1. Reuse Practices 17

2.4. Digital Humanities 18
2.4.1. Representations of Data 23

2.5. Situating Open Science Context Within Data Reuse 24

2.6. Data Discovery 28

2.7. Sharing Data 30
2.7.1. Disciplinary Differences Among Disciplines 32

2.7.2. Repositories and Data Communities 34

2.8. Data Curation as Best Practice 36
2.8.1. Data Curation in DH 38
2.8.2. Quantitative vs Qualitative Dichotomies 39

2.9. Data Librarians and Other Roles in RDM 39

CHAPTER 3: METHODOLOGY 44

3.1. Method 44
3.1.1. Interviews 44
3.1.2. Interview Schedule 45

3.2. Data Collection and Sampling Strategy 45

3.3. Data Analysis 47
3.3.1. Thematic Analysis 47
3.3.2. Coding and Themes 49
3.3.3. Limitations 54

3.4. Research Ethics 54
3.4.1. Risks to participants 55

CHAPTER 4: RESULTS 56
4.1. Results Overview 56
4.1.1. Participant Demographics 57
4.1.2. What Kind of Data do DH Scholars Use? 60
4.1.3. Participants' Disciplinary Domains 61

4.2. Structural Barriers to Data Reuse 62
4.2.1. Lack of Time and Resources 62
4.2.2. Accessing the Data 65
4.2.3. Inconsistent Data Sharing Practices 67
4.2.4. Dirty Data 69
4.2.5. Permissions to Access Data 71
4.2.6. Technical Training Gaps 74

4.3. Non-Structural Barriers to Data Reuse 75
4.3.1. Labour-intensity 76
4.3.2. Responsibility of Data Management 77
4.3.3. Qualitative, Quantitative and Mixed Data 80

4.4. Enablers of Data Reuse 81
4.4.1. Collaboration 82
4.4.2. Data Reuse in the Classroom 85
4.4.3. Technology, Tools and Machines 86
4.4.4. Curated Data Optimizes Data Reuse 89
4.4.5. Data Reuse Leads to More Reuse and Richer Data 91
4.4.6. Digital Humanists are Autodidacts 94
4.4.7. DH Scholars are Data Reusers 96
4.4.8. Data Reusers in DH have a Vision of the Future of Reuse 97

CHAPTER 5: DISCUSSION 99

5.1. Key Themes and Recommendations 99
5.1.1. Education 100
5.1.2. Shifting Roles in the Research Infrastructure 104
5.1.3. Data Stewardship, Curation and Long-term Data Preservation 108
5.1.4. Improvements to Data Discoverability 113
5.1.5. Data Sharing Pathways to Data Reuse 116
5.1.6. Culture Change 119

5.2. Further Research 123

CHAPTER 6: CONCLUSION 126

REFERENCES 133

CONFLICT OF INTEREST STATEMENT 163

APPENDIX A 164

APPENDIX B 165

APPENDIX C 166

APPENDIX D 169

LIST OF TABLES

Table 1: Research Data Management Roles ……..…….. p. 41

Table 2: Non-Structural Barriers to Data Reuse ……....p. 50

Table 3: Structural Barriers to Data Reuse.…….....p. 51

Table 4: Enablers of Data Reuse.…….....p. 52

Table 5: Types of Data Used…….……..…….. p. 60

Table 6: Disciplinary Background (after DH)…….……..…… p. 61

LIST OF FIGURES

Figure 1: Research Data Metaphor, Research Data Ecosystem..…p. 11

Figure 2: Research Data Metaphor, Data Curation Lifecycle …..….. p. 12

Figure 3: Research Data Metaphor, Three-legged stool……...….. p. 13

Figure 4: Participant Demographics, by region ……..……...….. p. 58

Figure 5: Career Stage…….……..…p. 59

Figure 6: Research Life Cycle and Data Management Life Cycle ..…p. 103

Figure 7: Strategy for Culture Change……..…..... p. 111

LIST OF APPENDICES

Appendix A: Sampling Technique and Results……..…. p. 154

Appendix B: Invitation Email for Recruitment of Participants…..…. p. 155

Appendix C: Letter of Informed Consent…….…….. p. 157

Appendix D: Interview Schedule……..……. p. 160

CHAPTER 1: INTRODUCTION

Scholarship is more and more data-driven, and as digital tools and the internet continue to evolve, sound data practices among scholars are now essential for scientific discovery (Borgman, 2015; Tenopir et al., 2020). Compared to the "recent age of computing", this present time can be more accurately characterized as the "age of data" (Manganelli et al., 2021, p. 1). In academia, data-intensive research has compelled changes that has affected science policy writ large, including but not limited to funder policies. An increasing amount of funding agencies, research organisations, institutions, data repositories and publishers are asking – and in some cases explicitly requiring – for data to be shared as openly as possible (Briney et al., 2015; Gregory et al., 2020a). In the United States, the White House Office of Science and Technology Policy declared 2023 as the Year of Open Science, announcing new actions to advance public access to research publications and data. Though the announcement focused on the hard sciences, the National Endowment for the Humanities Office specifically committed to funding DH projects, in an effort to address "issues of accessibility and usability, and designing equitable, open, replicable, and sustainable projects, including those that support the enhancement or design of digital infrastructure that contributes to and supports the humanities, such as open-source code, tools, or platforms" (The White House, 2023, para. 12). While institutions and funders laud the benefits of research data sharing and reuse, benefits can only be achieved if data are actually shared and reused by others (Borgman, 2015; Briney et al., 2015; Pasquetto et al., 2017). Research data

management (RDM) policies from public funders in Europe, Canada and the U.S. are requiring more RDM conditions to be met in order to obtain funding; publisher are increasingly asking authors to meet these conditions before being published in scholarly journals as well. Scholars who wish to obtain funding and publish their research therefore increasingly need to apply good RDM practices to the data collected for published research that meet funder and publisher conditions. RDM best practices often include data sharing and depositing (with appropriate access to the data where ethical, cultural, legal, and commercial requirements allow), and in accordance with funder, institutional policies, and FAIR principles.[1]

Data reuse is an essential component of open science (Pasquetto et al., 2017; Sielemann et al., 2020), but socio-cultural, technological, political, organisational, economic, and legal barriers stand in the way of opening up research practices like data reuse. Efficient, economical, responsible data practices have become meaningful to meet the goals of open science (Tenopir et al., 2020). Most recently, the Covid-19 global pandemic showed that there were life-saving benefits to immediate access to data and open access journal articles. However, while the pandemic may have caused an unprecedented increase in the number of open access publications, the publication of open data was not as widespread (Cerda-Cosme & Méndez, 2023). Reusing qualitative data from previously published projects can highlight older research, showcasing their value as new research questions are asked.

[1] FAIR: findable, accessible, interoperable, and reusable.

As the shift towards open science research grows, the present time is apt for examining data reuse as it settles its footing (Gregory et al., 2020b). The age of data is already here – a timely and relevant moment to study the qualitative, discipline-specific ways in which scholars work with reused data. Digital humanists are on the vanguard of data-intensive computational research, and consequently, it is important to better understand the data practices of researchers in a field that heavily uses data.

1.1. Disciplinarity and Data Reuse

There is a persisting view that humanities scholars do not work with data. Humanities scholars approach data collection differently than those in other disciplines – they are often contextualizing and analyzing the sourced material rather than merely collecting data and extracting features from them (Posner, 2015; Ruediger & MacDougall, 2023). Moreover, this fact is ambiguous in the Nelson memo, failing to distinguish between scientific data and sourced data (Ruediger & MacDougall, 2023). Nonetheless, it is helpful to review what *can* include humanities data collected for research:

> "Most of my colleagues in literary and cultural studies would not necessarily speak of their objects of study as 'data'. If you ask them what it is they are studying, they would rather speak of books, paintings, and movies; of drama and crime fiction, of still lives and action painting; of German expressionist movies and romantic comedy. They would mention Denis Diderot or Toni Morrison, Chardin or Jackson Pollock, Fritz Lang, or Diane

> Keaton. Maybe they would talk about what they are studying as texts, images, and sounds. But rarely would they consider their objects of study to be 'data.'" (Schöch, 2014, p. 2).

The humanities are unique in their approach and conceptualization to data for research. Increasingly, libraries and archives are adding to their collections of invaluable and unique research data and scholarly outputs, often of increasing size and complexity, all the while contributing to vast computational digital collections of different formats. Borgman (2015) has stated that data creation, use and reuse practices differ among academic disciplines, and further research into discipline-specific data reuse practices is needed. Data in the humanities is as much a "theoretical, methodological and social issue, as it is a technical issue", and DH aims to "realize the potential of this data for humanistic inquiry" (Schöch, 2014, p. 2). Siemens (2012) et al. argue that data reuse is foundational to the field of DH. It is an exciting, data-driven, and interdisciplinary field whose research centres on using computational methods to examine the cultural and heritage material of modern and ancient civilization. The choice to focus on DH in this study is inspired by its quandaries and engagement with Web history and archiving and new materialism. A new materialism approach to data reuse practices in DH can help us understand the motivations of its (re)users as entanglements between human and non-human matter (Hogan, 2015).

Qualitative research on data reuse in specific disciplines is scant, and the attitudes, practices and motivations for sharing and reusing data are therefore not well understood. There are few qualitative studies on disciplinary practices of data reuse,

notably, a 2017 DH study by Alex Poole. Poole (2017) used a cohort of DH grant recipients as a case study to look at the experience of American DH scholars' use and curation of research data, though he did not specifically concentrate on their experience of data reuse. Urszula Pawlicka-Deger (2021) takes a similar path as I do in this study – looking at open data in DH to situate it within a cultural research structure.

There have been a few ethnographic, qualitative studies about data reuse in the U.S.; among engineering scholars (Faniel and Jacobsen, 2010); among mostly computer science and engineering scholars in the Center for Embedded Networked Sensing study (CENS) from 2002 to 2012; and the DataFace Consortium study with mostly biomedical scholars from 2009-2019 (Pasquetto, Borgman & Wofford, 2019). The CENS study found that, unsurprisingly, the "degree of usefulness, trustworthiness, and value of the shared data" varies widely between researchers and disciplines (Wallis, Rolando & Borgman, 2013, p. 2).

While DH's preponderance to using new computational tools could help with data reuse, it may also hinder it, presenting its researchers with technical and policy challenges. As digital humanists re-interrogate old data using new tools, what might we learn about what is working well with data reuse, as well as what is not? By looking at how digital humanists use and reuse data, might there be some similarities among scholars who share datasets? Might repository managers, research experts and research librarians learn from and design better services from them?

1.2. Research Objective, Questions, Rationale and Goal

This study investigates data reuse by scholars in DH. It provides an analysis of interview data with participants to gain a better understanding of its challenges and opportunities in data use and reuse.

This thesis research study focusses on two research questions and is exploratory in its approach. **RQ1** – How do digital humanists use, find and reuse data? **RQ2** – What are the barriers and enablers to data reuse among digital humanists?

Data reuse is an understudied problem (Pasquetto et al., 2017). There have been calls for more research into the current practices, attitudes, and desires of researchers, especially in how they experience infrastructural challenges (Cooper & Springer, 2019) and attitudes towards open data policies (Baždarić et al., 2021). The goal of this study is to describe and thematically analyse DH scholars' experience with data reuse as a whole. The main intention is to understand how DH scholars use and reuse research data. The study's larger aim is to enhance current thinking and provide insight for information and library science practitioners at the intersection of DH and data reuse. Potential readers might be DH researchers, data librarians, discovery designers and repository managers from cultural heritage domains and academic institutions.

Potential readers of this research will be academic researchers, research support staff and librarians, data repository designers, government, and policy makers. In analysing the qualitative nuances of responses to data (re)use questions with DH

scholars, the goal of this research is to create a practical needs analysis of services for librarians working with researchers in DH.

1.3. Approach

This study uses semi-structured interviews. Qualitative interviews in research are used as a way to understand people in human–computer interaction research (Hwang et al., 2022). Semi-structured interviews are conversation-based (Hwang et al., 2022) "phenomenological encounters" between researcher and informant (Nowak & Haynes, 2018, p. 431). Phenomenology, the observation of social phenomenon, is particularly suited to qualitative research because it allows the researcher to take an embedded position(Creswell, 2014). Phenomenology seeks to describe the "essence" of a phenomenon by investigating it from the people who have lived it (Teherani et al., 2015). The approach to this qualitative study presupposes a constructivist learning framework, in which participants construct new knowledge through direct and active experience.

Semi-structured interviews are appropriate for studies in which research questions are exploratory in nature, as they are in this study. Exploratory questions help the researcher uncover unknown aspects of a particular topic – to understand "how and why things work as they do" (Gravlee, 2022, p. 69). There are disadvantages in conducting semi-structured interviews – some participants veered from questions asked, others talked for longer than others and therefore there was less time to complete questions in the interview schedule. And because questions were not asked

or answered systematically, depending on the time left in the interview, this made it difficult to compare responses between participants. However, when participants were less talkative, it was an opportunity to probe or re-ask the question in a different way. There was a balance of positive and negative aspects to this method, but the richness in detail provided by participants and the depth of conversation made it a worthwhile gamble. Most of my participants seemed comfortable being interviewed and leaned into the conversation about their craft. Semi-structured interviews as opposed to other phenomenological methods, is an appropriate method of inquiry because of its flexibility in format and its potential for the collection of rich data. Ultimately, the semi-structured method was chosen for this study because there is more flexibility than a structured interview.

CHAPTER 2: LITERATURE REVIEW

The literature for this research project came from several disciplines. Fields of study looking at data reuse have come from data and information science, and natural sciences (astronomy, health sciences, genomics). Since data reuse is an aspect of meta-research – an observation of the philosophical and practical underpinnings of research methods, reporting, reproducibility, evaluation, and incentives (Ioannidis, 2018) – all disciplines may have something to say about data reuse. This literature review is informed by information science and meta-research works and leans heavily on works by Christine L. Borgman in data scholarship, and Alex H. Poole in data in DH. Irene V. Pasquetto has published extensively on open data and disciplinary differences in data reuse. Bethany Nowviskie informs much of the future thinking in data scholarship and Thomas Padilla has published seminal work on Collections as Data. This review of the literature is divided into sections to examine the themes and issues. The main sections deal with researchers' use of data including data sharing, data discovery, open data, and open science, disciplinary and DH sharing and reuse, data communities, and roles for researchers and librarians in the data age. The review narrows its scope to an overview of data use contexts and practices to meet the goals of open science, the disciplinary and DH aspects of data reuse, and concludes with data discovery, sharing, and authorship.

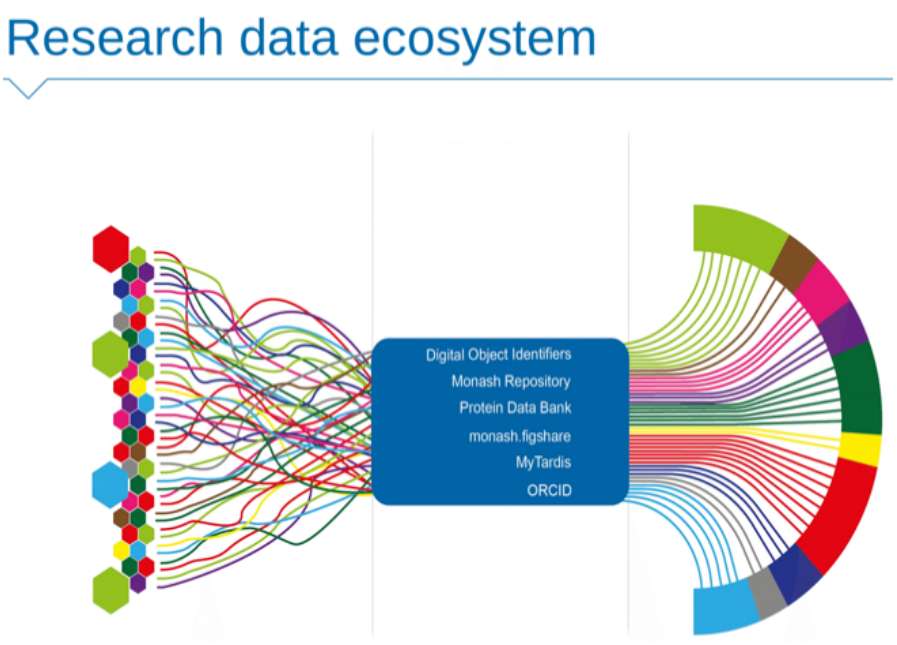
2.1. Defining Data and Data Reuse

Research data itself defies unilateral definitions: they have been referred to as "first-class intellectual objects" (Heidorn (2011, p. 671), ''scientific capital'' (Scaramozzino et al., 2012, p. 362), and "centrepieces of modern scientific discovery" (Tenopir et al. 2020, p. 2). Borgman offers its simplest definition: research data are "entities used as evidence of phenomena for the purposes of research or scholarship" (2015, p. 162). Research data – whether physical or digital – underpin scientific research findings and form the basis for scientific findings and outputs through peer-review, publication. The humanities and social sciences (HSS) have not agreed to one definition of research data. STEM fields use, reuse, and manage data very differently, and produce it at much larger scale than in HSS (Borgman, 2015). Defining what "counts" as data, is "a feat that remains frustratingly elusive" (Poole, 2017, p. 1772). For a DH scholar, data could be physical objects like 6000-year-old Indian coins, ticket stubs from a Shakespearean play in local township of 15th century England or the 1986 Women Writers Project, whose researchers have worked to digitize the London Times newspapers of 19th century England and augmenting the text corpora with machine-readable annotation. The Commons Open Repository Exchange (CORE) qualifies data as being digital objects like abstracts, articles, bibliographies, books, course material or learning objects, datasets, archival finding aids, interview transcripts, maps, sound and music, recorded performances, photographs, and visual art (Humanities Commons., n.d.).

Terms change as the technology evolves (Borgman, 2015), but metaphors can be useful because they are an example of how different scholars use different terms to mean the same thing. Current metaphors for research data systems include the *Research Data Ecosystem* (see Figure 1; Groenewegen, 2016), the *Data Curation Lifecycle* (Figure 2; Higgins, 2008), and even a *Three-legged stool* (Figure 3; Kenney & McGovern, 2003). The metaphors are interesting because they show the chaotic nature and variety of ways in which RDM is described in the current research culture.

Figure 1

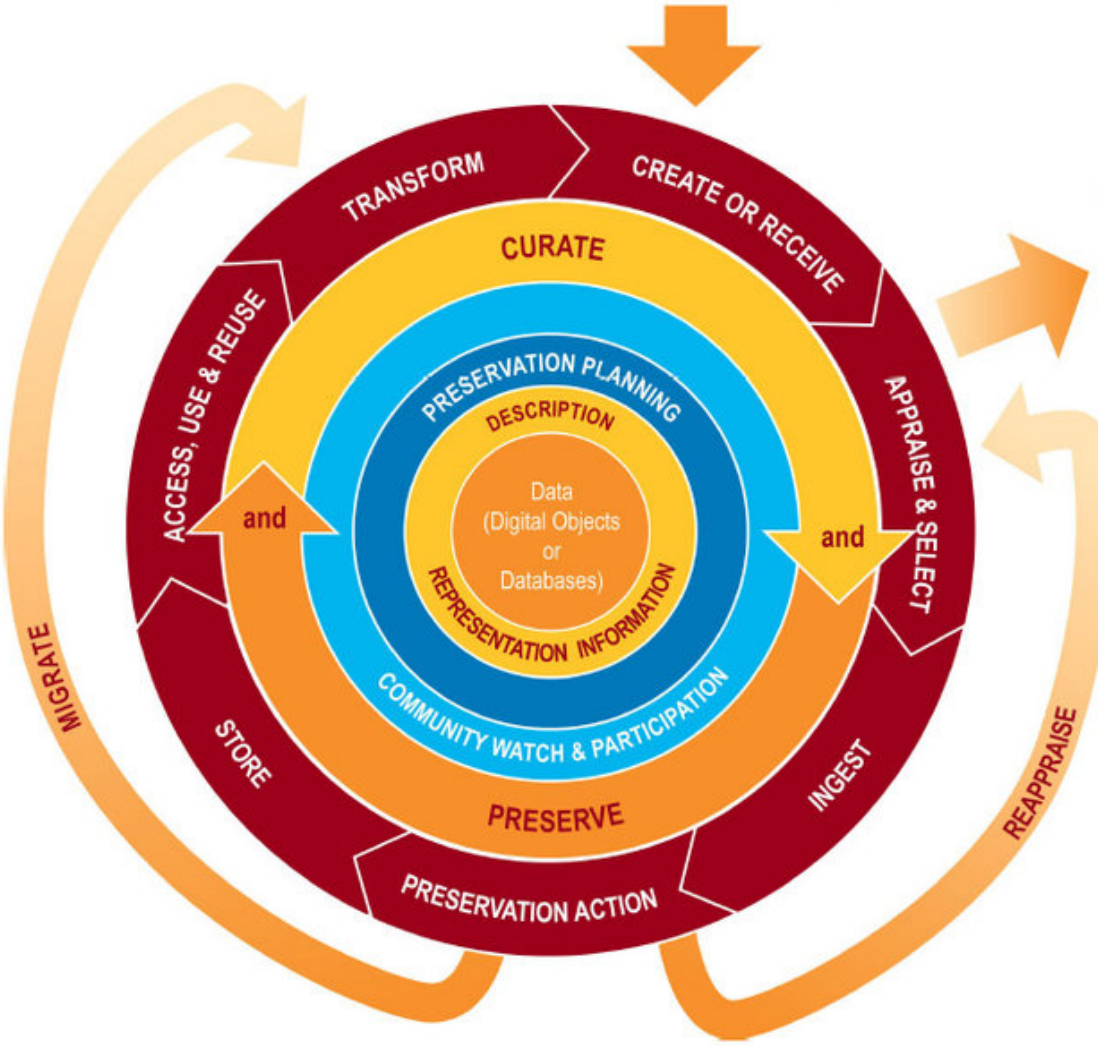
Research Data Metaphor: Research Data Ecosystem



Note. By David Groenewegen/Monash University Library. *Research Data Ecosystem* [Image]. From a slide in a presentation given at Figshare Fest in London, November 2015, https://bridges.monash.edu/articles/presentation/Research_Data_Ecosystem/3428336

Figure 2

Research Data Metaphor: Data Curation Lifecycle



Note. By Sarah Higgins. The macro to micro workflows within the DCC Curation Lifecycle Model [Image]. From The International Journal of Digital Curation, 2008, [image], https://blogs.helsinki.fi/thinkopen/know-your-data-rdm-series-1/

Figure 3

Research Data Metaphor: Three-legged stool

Note. By Anne R. Kenney & Nancy Y. McGovern. The Three-legged stool [Image]. From A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation?, RLG DigiNews, April 15, 2007, https://deepblue.lib.umich.edu/bitstream/handle/2027.42/60441/McGovern-Digital_Decade.html?sequence=4

2.2. Research Infrastructure and Data Scholarship

Borgman (2015) refers to data scholarship as a term to describe data practices beyond disciplinary methods. Data scholarship is also the data-intensive research practices scholars undergo in preparing for research projects to be published (Borgman, 2015). Information technology that supports research outputs has been called different things in different contexts to describe data intensive research since the early 2000s –

e-research, e-science, e-humanities (Borgman, 2015). Borgman uses the term data scholarship to more accurately describe data activities beyond disciplinarity, that is aligned with scholarly communication (Borgman, 2015). Data scholarship and scholarly communication help us understand the knowledge infrastructure of data-intensive research. Among those who reuse data, data use and reuse activities have various workflows and a wide range of components. Data scholarship can be difficult to define because there are many components in knowledge infrastructures – “robust networks of people, artifacts, and institutions” (Borgman et al., 2019, p. 2).

Research in data, open science and research infrastructure is nascent. The current research infrastructure favours open access – the publication aspect of open science. This has been successful, and changes are being seen in removing barriers in public access to scientific articles. The NIH for example has mandated that, starting in 2026, all publicly funded research publications be freely available immediately without embargo (Tollefson & Van Noorden, 2022).

Research funding often comes from government (e.g., Tri-Agencies, UKRI) or science philanthropy organizations (e.g., Chan Zuckerberg Initiative). And in the last several years, research data policies have been released by several government funding bodies internationally – European governments, the Canadian Tri-Agencies, the United States, and Australia. Many science philanthropy organizations (e.g., the Bill and Melinda Foundation or the Alfred P. Sloane Foundation) have specific conditions tied to grantees use of data after project completion as well. They differ in how they address the question of data after research is completed, but generally ask recipients to plan for

how they will manage their publicly funded data (usually through mandating Data Management Plans – DMPs). Some funders also have data deposit requirements – the Tri-Agencies have announced that they plan to require it in the next few years (no date is set yet). While they will not enforce a data *sharing* requirement, the data *deposit* requirement will be required, and is described in quite opaque terms: it is "encouraged" that researchers "provide appropriate access to the data where ethical, cultural, legal and commercial requirements allow" ("Tri-Agency Research Data Management Policy", 2021, para. 20). UK Research and Innovation (UKRI) state they do not prescribe *when* or *how* researchers should preserve and share data but does require them to make clear provision for sharing data when planning and executing research. UKRI believes that publicly funded research data "are a public good, produced in the public interest and that they should be openly available to the maximum extent possible" ("Data Sharing Policy", 2021, p. 3). To receive funding from Horizon Europe, researchers are required to submit DMPs, select appropriate repositories for long-term storage and share their data (Science Europe, 2020). Science Europe admits that this is often an administrative burden for scholars wishing to publish, and that requirements such as DMPs are viewed "as a bureaucratic burden, rather than a useful tool" (2020, p. 6). The variations in requirements may inspire likely feelings of "administrative burden" in researchers when thinking about data reuse and data sharing. There are more considerations to analyse in RDM policy frameworks globally, and the infrastructure questions to answer are larger than the space allowed in this thesis (Briney et al., 2005, provide a good analysis of U.S. RDM policies).

Major scientific journals who go forward with manuscript acceptance almost always have a data deposit condition for publication, including *Nature*, *PLoS*, and *Science*. Providing public access to scientific research is clearly a growing priority for public and private funders.

By and large, the research data infrastructure engages two cultures: the institutional or funder's policy administrative side, and the scholars' side (Pawlicka-Deger, 2021). The administrative side asks for data to be shared, reused, and opened and tie this demand to a national top-down requirement. It engages research infrastructure services such as IT, librarians, research support, information management, and RDM. And then there is the bottom-up, community culture engaged in the research data infrastructure: scholars who are required to manage and, for the most part, open research data. These two sides—administration and academics—are engaged with and motivated by different goals and outcomes (Pawlicka-Deger, 2021).

2.3. Overview of Data (Re)Use

What is meant by data reuse? In this research project, data reuse is defined as a "practice of using data by people other than the original collector" (Faniel et al., 2016, p. 1162). If data are collected by one individual, for a specific research question, that is its first use. It is considered a secondary use, or reuse, when a dataset is deposited into a repository, retrieved by another researcher, and then "deployed" for another project (Pasquetto et al., 2017). When I refer to data use and reuse, I use it in the context of research data for the purpose of knowledge production.

Research data use and reuse are complex constructs (Pasquetto et al., 2017). Reuse is not currently integrated into all research methods in academia, and for a variety of reasons: issues of data "authorship" (Duke & Porter, 2013); data misinterpretation (Niu & Hedstrom, 2008; Carlson, & Witt, 2010; Yoon & Kim, 2017); "data hoarding" (Strupler & Wilkinson, 2017); and the ethics and potential risks in data collected for research involving humans, notably in the health sciences (Frank et al, 2018). Research about data reuse in DH has only recently emerged. One study on Finnish newspapers showed that physical text reuse has been happening since at least 1771 (Salmi et al., 2021). One study even suggests physical text reuse was happening in 220 A.D. China (Sturgeon, 2018).

2.3.1. Reuse Practices

Wider adoption of data reuse among scholars in all disciplines may be lacking because of "dirty" data – incomplete, missing, or just poorly organized data (Niu & Hedstrom, 2008; Carlson, & Witt, 2010; Yoon & Kim, 2017; Gregory et al., 2020a; Hemphill et al., 2022). Inconsistent storage, archiving and preservation of data often prevents its reuse (Sielemann et al., 2020; Roche et al., 2015). There is often also ambiguity in authors' data availability statements[2] that accompany journal publications (Sielemann et al., 2020; Vasilevsky et al., 2017), and potential reusers may elect to skip

[2] Data availability statements usually include information about how the data were created, procured, where it can be accessed and whether it can be reused.

the dataset if they are uncertain of how to access the data. Data reuse benefits include decreased costs, time, and effort (Hedrick, 1998) – it is far more cost effective and time saving to collect data with reuse in mind (Goodman et al. 2014; Hedrick, 1988). It may lead to better, more transparent science as well, such as increased reproducibility of scientific findings and a better accountability of scientific research. Reusing data also have an impact on a larger scale – notably, coming up with new scientific discoveries (Sielemann, Hafner & Pucker, 2020; Safran, 2017; Pasquetto, Randles & Borgman, 2017). Under-utilised, yet reliable datasets lead to unnecessary data collection and may hinder scientific progress (Sielemann et al., 2020). Sielemann et al. also note that increases in data sharing may lead to higher quality, and a greater selection of, research data for reuse. The 2002-2012 CENS study found that data reuse was not consistently practiced among computer science and engineering scholars and that – unsurprisingly – the "degree of usefulness, trustworthiness, and value of the shared data" varies widely between researchers and disciplines (Wallis, Rolando & Borgman, 2013, p. 2). In the DataFace Consortium study, Pasquetto notes that it may be impossible to predict "how research data will be reused, by whom, and to what purposes" (2018, p. 232) and suggests that every reuse context is different. She concluded that collaboration, as well as well-developed metadata and ontologies documentation, was key to enabling data reuse among these scholars.

2.4. Digital Humanities

> "Digital humanities research combines humanities and social science research methodologies with computational techniques to allow processes such as data mining, text mining, data visualisation, data modelling, data analytics and text encoding" (Grant, 2016, p. 1).

Characterized as a dynamic, social, collaborative field, alongside what DH *does*, we must also deeply consider who digital humanists *are*. Nowviskie (2012) has stated that collaboration is a "hallmark" of DH – an aspect that sets the field apart. Its interdisciplinarity with other fields and its unique methods show that collaboration is part of and a condition of DH (Burdick et al., 2016). Burdick posits that DH goes beyond a characterization of humanities that does data – rather, it is a field that is marked by its ubiquity. The internet and the ability to digitized physical material in the humanities has opened up the networking and collaborative possibilities in data reuse. Burdick et al. call this phenomenon "ubiquitous scholarship". Ubiquitous scholarship has allowed the scholarly enterprise to expand via networks, information streams, and communities of practice that "produce and share knowledge and culture" which is marked by "an ethic of collaboration and interconnection" (Burdick et al., 2016, p. 59). DH, it may be inferred, is unique not only in its approach to data, but in its inherent collaboration.

It is just as important to define a field by who its practitioners are – who are digital humanists? Nowviskie provides a helpful list of who they could be: "…scholars working across disciplines, archivists and librarians, guardians and interpreters of cultural heritage, … technologists, developers, and specialists in method and form, … researchers, administrators, students, and shapers and makers of all kinds" (2015,

para. 5). Digital humanists also produce and often share the digital tools and code to analyze research data and material (Horik, 2019).

Relatively new as a discipline, the invention of the term “digital humanities” has only been in use since 2004. And research looking at the intersection of DH and data reuse is even newer and sparser. Up to now, research looking at discipline-specific data reuse has been limited to fields such as science in general (Kim & Yoon, 2017), engineering (Suhr, Dungl & Stocker 2020), STEM (Kim & Zhang, 2015), and ecology (Zimmerman, 2003). One approach has been to use information science theory to link digital curation practices to data reuse, such as the American scholar and archivist Alex Poole (2017). Poole’s research, like Gregory’s, touches on social interactions such as collaboration (2015, 2017; 2020b). These interactions in DH, Poole states, are central to data reuse (2015; 2017). The process of finding data to be reused is often mediated by social connections or personal networks (Poole, 2017; Gregory et al., 2020a). If access to a dataset a researcher wants to reuse is closed or unavailable, researchers often contact data creators for access. Poole suggests that collaboration is in fact more challenging in DH, more than other disciplines, because in DH there are such large differences in “language and terminology, methods and research styles, and theories, outputs, and values” (2017, p. 1776). Poole’s research intersects with DH, information studies and scientific data practices, as they are each a practice and a field of research. The areas of study share three elements in common. First, they are interdisciplinary and value collaborative work. Second, they study the science of science and methodology writ large in academia. And finally. The work of practitioners in these domains often

work both inside and outside academic institutions (Poole, 2017). DH, information studies and scientific data practices also intersect in their challenges, for example in their “sustainability, project management, institutional positions, and scholarly valuation of their work” (Poole, 2017, p. 1772).

DH may also be considered as a method, by nature of its emphasis on using new and emerging computational tools to explore or re-examine old humanities corpora and collections. DH is a wide and amorphous field of study beyond methodology. The field is awash with existential questions about the essence of DH – are you still a digital humanist if you don’t “make” anything, computationally? If you’re not using new computational tools to answer old questions, is it still DH? The question of more data often becomes a logistical one – the many ways that DH scholars will find, store, use, share and reuse data. It is a natural debate that helps give the field its excitement, its relevance, and proof that the borders of this discipline are amorphous, its methods diverse and fluid. There is a multiplicity of cross disciplinary approaches to DH. Just to name a few: Safiya U. Noble’s work in Critical Internet inquiries, Jacqueline Wernimont’s work in DH and Intersectional Feminism, Bethany Nowviskie’s research in human-machine artifacts, or Jada Watson’s work examining the country music industry’s devaluation of BIPOC artists.

Scholars in DH use a variety of physical and digital objects or artifacts for research and analysis. Digital objects in DH are data representing cultural or heritage material that have been digitized, catalogued, and often indexed and preserved for the short or long term. They are often stored as web archives – collections on the websites

of museums, university repositories or archives, cultural centres and local archives, or digital repositories. The web contains an enormous amount of data desirable for research by digital humanists, and Milligan suggests that history research is being reshaped by the web (Milligan, 2022). To give an idea of scale of data stored on the web, there are over 778 billion URLS that the Internet Archive (IA) links to, preserving about 42 PetaBytes of books, music, videos (Internet Archive: Petabox, n.d.). The IA stores and preserves much of the web's data and is just one of several data sources used for data discovery by scholars in HSS. Their collections encompass digitized material such as photos and scans of physical objects, massive textual corpora, and web archived content such as discussion forums, blog posts, obsolete journals, manuscripts, monographs, and websites captured by the Wayback Machine. The IA is one example of large-scale DH material in preservation.

In DH research, the field has started to take a "more data-focused look" towards cultural heritage collections and data literacy, notably with Thomas Padilla's "Collections as Data" project (2016). Padilla describes a cultural shift where digital humanists have started to see the potential for GLAM material to be treated as whole data collections. Collections as Data is also a call to action in the field – an invitation to merge the world of big data with GLAM through the use of new computational analysis tools like text mining, data visualization, mapping, image analysis, audio analysis, and network analysis. Seeing cultural heritage collections as data allows DH to collectively approach collections as individual "wholes", and act as stewards of these corpora: "Collections as data raises the question of what it might mean to treat digitized and born digital

collections as data rather than simple surrogates of physical objects or static representations of digital experience" (Padilla, 2018, p. 296). The project calls for a reconceptualization of GLAM collections as valuable research data collections. The Collections as Data approach comes with similar philosophy as the open science movement and is aligned with values seen in librarianship – reflected in three of the 10 principles in The Santa Barbara Statement: "Collections as Data should be made openly accessible by default, (except in cases where ethical or legal obligations preclude it), should be interoperable, and its proponents are stewards that should "work transparently in order to develop trustworthy, long-lived collections" (Padilla et al., 2019, p. 296). The project is a good indicator that DH has a tradition of working closely with data.

One of the barriers that scholars in the humanities have faced in recorded history, is that humanities data is often so "rich and complex, non-standardised in format" yet "without common or consistent metadata and ontologies." (Horik, 2019, p. 5). This is where representations of data can be helpful.

2.4.1. Representations of Data

The humanities have long struggled to find ways to represent data for "input, processing, and output" (Horik, 2019, p. 5). In recent years though, DH developed its own meta-data schema to communicate with other DH scholars – the Text Encoding Initiative (TEI). TEI is a standard that encodes structural and linguistic elements of a text in machine-readable way, using TEI to represent texts for scholarly editing or analysis.

This mark-up convention has been around since the 1980s, but scholars have rediscovered its usefulness in about 2004 – it is good at making cultural heritage annotation text machine readable. It also saves time by serving as a publishing and preservation tool for DH data (Flanders & Hamlin, 2013).

2.5. Situating Open Science Context Within Data Reuse

The push towards better data use and reuse practices from funding agencies, research organisations, institutions, data repositories and publishers stem largely from a global movement that is slowly changing the culture of science publication. Open science is a movement and a paradigm shift in how the methods of science are conducted, enabled by technology, and marked by collaboration (Ramachandran et al., 2021). Proponents of open science have long advocated for data practices that help research align itself FAIR Principles (GO FAIR Implementation Networks, n.d.). Conversely, closed science and poor digital research infrastructure, combined with low data literacy, threatens open science (Tenopir et al., 2020). The benefits to data reuse are numerous and have been well-documented in published studies from multiple disciplinary perspectives. Reuse of data is a marker of the transition from closed to open science, is essential to the goals of open science and holds “immense potential for scientific discovery” (Sielemann, 2020, p. 7). Some of the most oft-cited benefits of data reuse are the maximization of time, labour and cost, the expansion of scientific knowledge, and positive impacts to authors (Sielemann, 2020). Data reuse helps “de-silo” academia, prevents duplication, and reduces research waste (Hogan, 2015; Purgar

et al., 2022). In this age of data, there are economic gains to be had in data reuse – funders and institutions are eager to maximize efficiency in research outputs. Researchers using data collected by others may benefit from saving time, cost, and resources. Although it is difficult to put a number on the monetary value, data reuse is intrinsically tied to a data economy – a “value creation proposition” tied to data-intensive infrastructures (Tempini, 2017). The data reuse era is already here, and it is important to understand the qualitative, discipline-specific ways in which scholars work with reused data. It is especially important to closely examine and understand the data practices of scholars who are on the vanguard of data-intensive computational research such as digital humanists.

Terms such as “open data,” “open science” and “open research,” are poorly understood (Pasquetto et al., 2015). The notion of “open” in research has been developing since the 1950s and one of the earliest disciplines to adopt open data was in the 1980s with trends within natural sciences, in genomics (Mauthner & Parry, 2013). Open also has strong roots in the free-software movement of the1980s, followed by the open source and open content movement. Open source is shared, editable software or programming code usually meant for web developers, and open content is licenced, or public domain content used to populate knowledge systems such as the free online encyclopaedia, Wikipedia. Open has, in recent years, has also included open educational resources, open data, open government, open knowledge, open access, and open science. Open Science includes open data outputs as contributions to the knowledge infrastructure, and Open Access refers to the publishing of scholarly journal

articles without restrictive access. Open Scholarship is becoming the all-encompassing, if extremely broad, term to describe a number of open practices within academia (*UBC Wiki*, n.d.). Open access and open data in Canada emerged in the early 2000s, and in 2020 the government proposed a Roadmap for Open Science (Office of the Chief Science Advisor of Canada, 2020). In European countries, this concept emerged earlier on, with global declarations. Some of the first declarations and white papers co-signed by members of academic communities around the world included the Budapest Open Access Initiative (2002), the Bethesda Statement on Open Access Publishing (2003), and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003).

There are disciplinary challenges in open data practices (Curty et al. 2017; Pasquetto et al., 2016; Poole & Garwood, 2020; Yoon & Kim, 2017), but this thesis does not look closely at the motivations around open data practices. It is important to note that there is a wide spectrum of “openness” in open data (Open Knowledge Foundation, 2020; Ramachandran, 2021). The main themes around reasons for making data open often comes back to the perceived risks around intellectual property (IP), privacy and even spreading misinformation. As well, scholars’ have concerns about rewards and are often disincentivized to share and publish datasets, which may lead to data hoarding (Hogan, 2015). If members of the general public gain access to the data without scientific expertise, there be risks in misunderstanding or decontextualizing the data – the real possibility of nefarious actors misusing research data to spread misinformation, or publish incorrect or misleading outputs (Edmiston et al., 2021; Strupler & Wilkinson,

2017). Some of the advantages of open data sharing and publishing are the potential for increases in citation metrics for scholars. It is also a question of doing “better” science – more transparency in methods and processes to outcomes, the ability to reproduce results, and to encourage community engagement in knowledge creation and evidence building (Piwowar & Vision, 2013; Sielemann et al., 2020; Seastedt et al., 2022; Stieglitz et al., 2020; Strupler & Wilkinson, 2017). Further, open data that are scraped from government websites or databases also have concerns, such as the long-term accessibility of such datasets and whether replication will always be possible. Conversely, critics say that making research data openly accessible has personal and reputational risks for researchers including data theft, loss of IP, legal or administrative reprisal, and tarnished reputation (“Safeguarding Your Research”, 2021).

Open science in regard to data reuse does not follow a linear path. While there is a clearer line between open science and data sharing, discussions around data reuse and open science are often obfuscated – ethical considerations of open data often muddle the issue. Better training to increase data literacy, knowledge and understanding of open science and the “standard, tools, processing pipelines, and protocols” of FAIR are indispensable as research moves towards becoming more open (Ferrari et al., 2018, p.48). The attitudes, practices and motivations for sharing and reusing data need to be better understood in the humanities, and certainly within each sub-discipline. Though data reuse is not yet a normative and consistent practice for all scholars, RDM education and training to scholars – with a disciplinary approach to the

training material – could potentially lead to more amenability to adopting open science and data reuse practices (Sielemann et al., 2020; Wang et al., 2021).

When researchers practice open science by collecting, creating, transferring, and re-using scientific material, it can be argued that one of the positive outcomes of open sciences is increased collaboration within the knowledge infrastructure (Organisation for Economic Co-operation and Development [OECD], 2020). If science is to become more collaborative, one pathway is though the management, sharing and reuse of data (Faniel & Jacobsen, 2010). Yet open science requires cultural change – an overhaul of the knowledge infrastructure – "in order to build a new model of scholarly interaction and collaboration" (Pawlicka-Deger, 2021, p. 41). Socio-technical practices such as data reuse are social and technical enablers that facilitate knowledge sharing (Choi et al., 2008). Data reuse as a socio-technical practice therefore cannot be detached from its larger system – knowledge infrastructures, knitted together by organization, technological and cultural elements (Choi et al., 2008; Gregory et al., 2020a; Pandey & Dutta, 2013).

2.6. Data Discovery

There are legal, technological, societal, human, and economic barriers that inhibit the use and reuse of data (Bachlechner & Custers, 2017). It is important to understand data search as a socio-technical practice. Socio-technical research's goal is not to examine the social and the technical in isolation, but to look at "the interactions that occur where the two intersect" Gregory et. al., 2019, p. 473). Data discovery is a

necessary mechanism to data reuse, as Gregory shows in a typology of the information-seeking practices of data reusers (Gregory et. al., 2020a). Gregory et al. found that data reuse actually requires several conditions before it is possible, for example, having some form of interaction or collaboration between the data reuser and the data creator.

Discovery of data suitable for reuse is sometimes impeded by search difficulties and inadequate indexing (Willinsky & Wolfson, 2001). The data discovery experience is mediated by online search environments like data repositories, university library discovery systems and search engines. Users prefer to use discovery systems they know and are comfortable navigating. Other data discovery methods may include using search strategies like citation-chaining or serendipitous data discovery (Gregory et al., 2020b), or social ways such as hearing about a dataset through a colleague or at a conference. Human interaction in data search and discovery is seen most often in settings where interests and areas of study co-mingle (Gregory et. al., 2020b). Domain repositories can help connect scholars to each other, share and discover discipline-related data, which then may enrich the user's experience and help build data communities (more on this in 2.5.2. Repositories and Data Communities).

Information scientists have conducted important research into the data seeking behaviours of data users and reusers (Wang et al., 2021; Gregory et. al., 2020b; and Yoon, 2017). All kinds of people search for research data. A data seeker or potential data reusers could be a researcher, a librarian, a data scientist, a citizen scientist, or a machine – each possess unique search and discovery heuristics (Gregory et al., 2020a). Scholars face technical barriers to data reuse, especially in the exploratory,

searching and discovery phases of research. Scholars are more likely to use datasets that are easy to understand, well-organized, interoperable and tend to focus their search efforts on gaining access to data that easy to find and use (Gregory et al., 2020a). Scholarly data reusers tend to critically examine the data to determine whether to reuse it or reject it (Faniel, Kriesberg & Yakel, 2021; Faniel & Jacobsen, 2010; Rolland & Lee, 2013; Zimmerman, 2008

Data discovery and data seeking behaviour is inseparable from the technical aspects of data reuse, because “data search is a complex phenomenon grounded in the interplay between technology and social practices, but not reducible to either” (Gregory et. al., 2020b, p. 473). Data reuse is inherently a socio-technical practice because data reusers often build their work on other researchers' findings (Gregory et. al., 2020b). Better design of data discovery infrastructure is key to meeting data seeker’s needs (Gregory et. al., 2020a), and well-indexed domain repositories enrich the user’s experience and help build data communities.

2.7. SHARING DATA

The economics of data reuse is multi-layered and complex. There are clear barriers to reuse, but chief among them are inconsistent data sharing practices, and even when data is shared, difficulties in discovering data (Gregory et. al., 2020b). Data sharing may in fact even be *the* condition that needs to be met *in order to* reuse data. Scholars tend to be better at following data sharing policies when there are personal benefits to their career as academics (Borgman, 2015). As the “enabling” practice to

data reuse, data sharing may also be central to open science since reuse is conditional on this research data management practice (Tenopir et al., 2020).

Data sharing has been an ongoing concern in humanities since at least 1988. Hedrick was one of the first to remark that HSS scholars should share their data, pointing out a number of factors impeding data sharing among them – for example cost, anxiety about losing control of the data, confidentiality, time and effort, the risk of poor reanalyses, a lack of institutional support, and the risk of being scooped (Hedrick, 1998). Yoon and Kim (2017) suggest that these barriers exist because of uncertainty on the part of the researcher in how their research materials will be used once shared. In other cases, copyright and permissions may not be clearly stated and will complicate its reuse (Horik, 1988). In fact, some researchers are unsure how to identify data ownership, or data reuse permissions and conditions (Briney et al., 2015; Murray-Rust, 2010). Datasets that are derived or "sourced" from one or more databases may also pose a challenge because scholars are unsure if it is possible to reuse in-copyright material (Grant, 2016).

The frequency with which data are shared, or reused, in any discipline, is "extremely difficult" to track (Pasquetto et al., 2017, p. 2-3). This is not only because these concepts mean different things to different scholars, but because there are so many contexts in which data sharing can occur (Pasquetto et al., 2017). Instead, sharing and reuse metrics have been largely based on self-reporting statistics, up to now (Pasquetto et al., 2017). Studies show relatively low levels of data sharing and reuse (Faniel & Jacobsen, 2010; Pasquetto et al., 2020; Treadway et al., 2016; Wallis et

al., 2013). Researcher do seem to have the intention to share data – there are higher rates of "intended" sharing and reuse (compared to actual) (Pasquetto et al., 2020; Treadway et al. 2016; Tenopir et al. 2011). The journal *Science* showed low rates of data deposit, reporting survey results querying the journal's peer reviewers over ten years ago – 8 % deposited their data into a repository, and 88 % stored their data on university servers (making it unavailable to researchers outside the institution) (Science Staff, 2011; Borgman, 2015). A 2021 faculty survey in the U.S. found that less than 20% of humanities faculty members reported sharing or depositing research data (Blankstein, 2022; Ruediger & MacDougall, 2023). At a conference hosted by the Journal of Open Humanities Data, only 47% of attendees informally surveyed said they had previously deposited research data into a repository (Ruediger & MacDougall, 2023).

There appears to be more research about data sharing in the natural sciences than the humanities, suggesting that data sharing in natural sciences is becoming more and more normative there (Sielemann et al., 2020). Data reuse, however, still has a way to go before it is as normative, consistent and a widely adopted practice (Sieleman et al., 2020).

2.7.1. Disciplinary Differences Among Disciplines

It is interesting to note the arguments against data reuse, and data sharing, in the health sciences only to compare the disciplinary differences between the HSS and STEM. There is disparity and little agreement among researchers about how, when and

in which context to reuse research. Some disciplines, like health science, have long discussed the ethics of data reuse and sharing. Ensuring that privacy is upheld is one of those areas of concern among the medical community, especially in medical imagery (Savage & Vickers, 2009). Scholars in health science often warn of risks to patient privacy – if disaggregated, identifiable patient information is openly shared, there is a risk of harm to the patient if their data is shared or reused without prior appropriate context and informed consent. Some scholars argue that deidentification is “a myth” (Lotan, et al., 2020). Pasquetto (2018) cautions that the rights of data donors (patients, study participants) must be respected since they have little to no control over how open research data are reused. It is difficult to manage or regulate how datasets will be reused once deposited to an open repository, however, and data reuse can cause “unintended consequences” (p. 177). But while the risk to patients should be minimized, “this risk will never be zero” (Seastedt et al., 2022, p. 10) and society should decide when and where data sharing should occur for the benefit of the knowledge infrastructure. Data sharing is complex, and there seems to be few incentives to share in most disciplines. Rowhani-Farid, Allen, and Barnett (2017) suggest that low rates of data sharing in biomedical health are due to low-rewards or no incentives in data sharing for the scholar’s career.

Yoon and Kim state that RDM training among scholars ultimately helps strengthen the knowledge infrastructure: “scientists without data reuse experience may have more concerns about misuse of data. As those concerns can be an important impediment to reusing data, it is important to provide proper education and institutional

support to address such concerns" (2017, p. 231). There might also be slight disciplinary differences within DH. The sub-disciplines that come together under the humanities (for example, history, anthropology, language, visual and performance arts, etc.) combined with the computational aspect of DH means that a DH scholar approaches and values "data" differently compared to STEM. Borgman notes that the humanities are actually the least likely of the disciplines to generate their own data (for example, STEM scholars' data often includes observations, models, experiments); they will instead rely more heavily on records of human experience as data sources such as maps, newspapers, photographs, correspondence; birth, marriage, death, and school records (2010). One of the major difficulties among DH, and humanities at large, is their dependence on sources of data often owned by others, and lack of control over the intellectual property rights of sources they use in their analyses. Comparatively, STEM scholars tend to work with their own IP – original observations or specimens (Borgman, 2010).

2.7.2. Repositories and Data Communities

Another way for DH scholars to discover research data would be through an institutional repository affiliated with a university of college, a generalist repository like Dryad, Dataverse, Figshare, Zenodo, or a domain, discipline-specific repository. A repository is the technical mechanism through which data reuse is made possible. There are advantages and disadvantages to using any repository. Some repositories are free to use, some cost have a cost associated to them. Some repositories have

restrictions such as requiring depositors that have an institutional affiliation, and others do not. Some repositories only host searchable metadata due to the sensitive nature of the data (private or restricted) and are usually obtainable under conditions, and on a case-by-case basis. There are several institutional repositories indexed by Google Scholar ("Google Scholar: Metrics," n.d.). In Canada, the Canadian Association of Research Libraries (CARL) estimates there are 94 institutional repositories hosted by universities, colleges, and non-profits (CARL, 2018). Institutional repositories do not index each other, nor could they, given all the working languages of universities in the world.

Aside from institutional repositories, commercial repositories also store research data, for example Kaggle, Google Cloud Public Datasets, for-profit cloud repository hosting service GitHub, and many more in many countries and languages. There are 128 million repositories on GitHub (Kashyap, 2020), yet for scholars, institutional repositories may be seen as more trustworthy than other options (Hemphill et al., 2022). In Canada, CARL cites seven general-purpose repositories – which accept data from researchers without access to an institutional repository. In the U.S., most domain repositories funding goes to the natural sciences rather than the humanities (Ruediger & MacDougall, 2023), though the earmarked funding for National Endowment for the Humanities announced by the White House Office of Science and Technology Policy (OSTP) in 2023 may help bring equitable funding to humanities repositories.

Cooper and Springer (2019) argue that domain repositories can create and unite data communities. Domain repositories may in fact "play a critical role in facilitating data

sharing and reuse" and could serve as sites of collaboration for "data communities—fluid, interdisciplinary, networks of scholars with overlapping research interests" (Ruediger & MacDougall, 2023, para. 5). Data communities could play an essential role in "fostering vigorous, voluntary sharing and reuse of research data in the sciences" (Ruediger & MacDougall, 2023, para. 5).

It can be difficult to select the appropriate repository, and scholars often need to consult with institutional librarians to find the appropriate repository. Although academic librarians can be good resource for scholars to consult with, such as finding a repository that will accommodate the size of the dataset or which format of files are supported, there are other options. Domain repositories tend to make this process easier because of their domain expertise (Buddenbohm et al., 2021; Habermann, 2023). Services like the global registry of research data repositories service, Re3data and the Data Deposit Recommendation Service, for European humanities scholars (Buddenbohm et al., 2021) are also helpful to scholars without data services support.

Borgman et al., in a case study of DANS (a Dutch repository for research data), observe that a good repository is part of a larger knowledge infrastructure, and when it works well, the research data reuse and sharing arm of this infrastructure should "mediate access to data by providing human, technical, and policy infrastructure for their communities" (Lee et al., 2006; Borgman et al., 2019).

2.8. Data Curation as Best Practice

One of the themes in the literature addressing barriers in data reuse is data curation. Curation, which means “to care” in Latin, is a process that adds value, maximizes access, and ensures long-term preservation to data (ICPSR, n.d.). Data curation is akin to work performed by an art or museum curator. Through the curation process, data are organized, described, cleaned, enhanced, and preserved for public use, much like the work done on paintings or rare books to make them accessible to the public now and in the future. The Committee on Data of the International Science Council (CODATA) defines curation as “a managed process, throughout the data lifecycle, by which data/data collections are cleansed, documented, standardised, formatted and inter-related” (“Data curation”, 2021, para. 1). Without curation, however, data can be difficult to find, use, and interpret. Hemphill et al. (2022) notes that there is a correlation with curation and reuse – they are downloaded more often than uncurated datasets or data collections. As well, uncurated datasets tend to be “minimally accessible” (Hemphill et al., 2022).

There are many goals of data curation, but mainly its aim is to manage the use of data from its point of creation throughout the lifecycle until it is reused for new research. Good curation practices ensure data are fit for use and reuse, and discoverable. Well curated data in the research infrastructure makes research data findable, accessible, interoperable, and reusable. Curation contextualizes and optimizes metadata standards and schemas to help data be discoverable. Repositories host research data, providing online spaces with technical capacity for describing and storing data, preserved on servers all over the world. These may be housed in academic or heritage institutions , or

in commercial digital spaces. Repositories engage in a practice of indexing in order for the data to be discovered, and eventually reused. Curating data is becoming the best practice in promoting and optimizing data reuse (Dekker & Riegelman, 2020; OCUL Data Community Data Rescue Group, 2020; Poole, 2015) and well-curated data helps scholars discover data available for reuse. There are DH-specific ideas around data curation, which I describe in more detail in section 2.3.3.

2.8.1. Data Curation in DH

Data curation is not well understood in the humanities (Poole, 2017). Poole's qualitative studies on data curation practices of digital humanists in the U.S. found that their challenges in data curation have implications for the future use and reuse of data. In interviews, participants described difficulties in managing their own research data, especially in regard to data curation activities like archiving, preserving, describing their own data, as well as challenges in collaborating and communicating, project managing, resource managing, and troubleshooting various technology. Poole's interviews with DH scholars showed that although they were interested in data curation training, they overwhelmingly faced barriers to identifying their training needs and gaining this knowledge because of various communication difficulties – languages spoken, terminology used in DH methodology, and varying levels and practices of data management (2017). Research in the computational side of DH does not look at curation extensively, and this lack of research points to a gap in training for general data management skills (Grant, 2016).

2.8.2. Quantitative vs Qualitative Dichotomies

DH, as a discipline, uses mixed methods, and does not appear to be tied to any particular direction either way. In fact, any optic of collecting either qualitative or quantitative may in fact be a “false dichotomy” in DH (Walsh, 2012). One of the most unique features of data reuse in DH, is where a researcher attempts to reuse a quantitative dataset and transform it into a qualitative dataset (Spiro, 2012). DH is a pedagogical practice that “pivots on learning by doing and by producing, blending quantitative and qualitative methods” (Garwood & Poole, 2018, p. 550).

There are some interesting examples of data reuse in practice. Jane Gray transformed quantitative research into a qualitative by interviewing 113 Irish children that had been surveyed in the 20th century. Drawing on life-history narratives, Gray used the new data (qualitative life stories and memories told in their old age) alongside the reused data (quantitative survey answers) and created a new research output. In another example, Wiggins (2015) reused qualitative interview transcriptions from Ann Oakley’s 1979 *Becoming a Mother* project. After interviewing the same participants from 35 years earlier and using computational tools to compare and analyse the new data with old transcription data, Wiggins published a new scientific output (Bishop & Kuula-Luumi, 2017). Sukumar and (2019) Metoyer warn that qualitative research does not naturally lend itself to data reuse, because it can be harder to replicate due to being specific to a time and a place.

2.9. Data Librarians and Other Roles in RDM

Within librarianship, there are new ways of exploring the role of librarians within the landscape of the data reuse paradigm, and Information science literature on data reuse centres the role of the librarian. Alongside discussion of library data service offerings for researchers, librarians are referred to as experts – professionals who are perceived as having expertise and advanced skills in data searching and discovery (Jeffery, 1998; Poole & Garwood, 2018; Nitecki & Davis, 2019; Ohaji et al., 2019). There are currently four types of roles (data steward, data governance, data librarians, information managers) for people working with research data and scholars (see Table 1). Contrariwise, an academic library (and by turn its librarians) is "not always the appropriate scale on which to address the challenges that scientists face" (Cooper & Springer, 2019, p. 24). Nonetheless, librarians with a domain specialization who wish to take on the challenge should become data curators that champion data sharing either from within inside their home institutions, or via data curation networks (Cooper & Springer, 2019). This is considered a bottom-up, community approach, where it is much harder to affect change at the infrastructure level (Cooper & Springer, 2019). Community-scoped, bottom-up approach in which care is taken to examine and address scholars' practices at the community level may be what is needed to achieve systemic change (Cooper & Springer, 2019).

Table 1
Data Management Roles

Role	Description
Data Steward	A data steward oversees each aspect of the data lifecycle, defining operational procedures to meet the requirements stated by organizational policies regarding the creation, collection, storage, or use of, and denial of access or data. (Statistics Canada, 2020).
Data governance	An overall organizational activity, currently in need of implementation and culture change. Data governance is strategic, high level and requires senior leadership of institutions to create and nurture the organizational structure required to sustain good RDM (Statistics Canada, 2020).
Data Librarians	Still a nascent field of professional work, data librarians' role is to support

	researchers' data needs and are other either data generalists or subject specialists in a discipline where this knowledge is warranted (Federer, 2018).
Information Managers	Responsible for the capture, digitization, representation, organization, transformation, and presentation of information ("Computer Science - Information Management", n.d.)

Ohaji, Chawner and Yoong (2019) propose a framework for expanding the role of data librarians in order to enable open science. The authors emphasise the importance of libraries in recruiting, training, and retaining librarians with professional data services competencies. Librarians can help meet researchers' data needs by lending their knowledge and expertise in data curation, search strategies, accessing training to learn new technology, information and project management, publishing, and metadata standardisation (Ohaji et al., 2019). There are challenges in allocating resources for data services, and libraries may not be prepared to provide a data service because of competing organizational, national, international context and infrastructure needs (Ohaji et al., 2019). Poole and Garwood assert that librarians should be considered as partners in research, but are currently being underutilized (Poole & Garwood, 2017).

Data curation can be an entry point for librarians who want to connect with DH researchers, and it in fact may be their most “robust opportunity” to work with digital humanists (Poole & Garwood, 2017, p. 818). Other interventions for librarians exist for those who wish to work with these scholars beyond curation, for example helping with data discovery, improving citation metrics by linking to persistent identifiers, providing guidance and access to data deposit platforms, tools for project management, recommendations for data sustainability among storage platforms, as well as offering training to support learning in curation and RDM (Poole & Garwood, 2017).

CHAPTER 3: METHODOLOGY

This chapter lists the methods, instrument, and research design of the research study.

3.1. METHOD

This study used semi-structured interviews with 12 participants selected at random from presenters of an annual, international DH conference3, ADHO. All twelve interviews were conducted in winter 2022. The instrument was, in line with phenomenological research, the researcher. The analysis was conducted using Thematic Analysis (TA).

3.1.1. INTERVIEWS

The interview as a data collection method lends itself well to this research project because it offers an opportunity to detect subtleties in the experience and worldviews of participants. Interviews offer the value of spontaneous, unfiltered reflections from participants. Observing social cues like facial expressions and voice intonation can help the interviewer control and steer the conversation back towards the research questions (Opdenakker, 2006). The semi-structured interview format is also a chance for narrative. DH researchers spoke about data use/reuse challenges and opportunities in a somewhat free and informal manner, with time to reflect and the opportunity to return to previous questions. Conducting interviews also has the potential for participants to

[3] https://adho.org/conference

provide deeper responses than other qualitative methods (Wildemuth, 2017). The element of human interaction in interviews enables both the participant and researcher to clarify, return to or re-ask questions where there is ambiguity or uncertainty. Semi-structured interviews in particular offer less rigidity and more leeway than the structured interview, while also being more organized and systematic than the unstructured interview (Wildemuth, 2017).

3.1.2. Interview Schedule

Each interview took place through the Microsoft Teams video conferencing platform with meeting links sent through an email invitation. Participants were asked to return signed consent forms by email before the start of the interview. Interviews were planned for 60 minutes in duration, with most interviews taking 45 to 70 minutes. Interview questions were grouped under three thematic blocks: Data use and reuse (Block I); Research data management (Block II); and Data sharing and reuse (Block III). A fourth block of questions included the closing question and if time permitted, an open-ended question about their opinion of DH and data reuse (see Appendix D for interview schedule and script).

3.2. Data Collection and Sampling Strategy

Random sampling. The population under study for this project were individuals who work and identify as DH scholars or researchers. Participants were drawn from a targeted population which was pulled from three years of conference programming material between 2018-2020. This information was found on each annual conference

website designed each year, which lists every speaker and paper presented. The conferences in the sampling were the following: ADHO 2018 at *Universidad Nacional Autónoma de México*, ADHO 2019 at *Universiteit Utrecht*, and ADHO 2020 at Carleton University and University of Ottawa. ADHO was selected because it preserves more than 20 years of conference material. It is also known for being the longest-running international DH conference. Conferences have been running since 2006 under its current name, but began running under another name in 1989. As one of the core gatherings in the field of DH, it is for thought of as “the one event in the academic calendar at which their research can be appreciated in all its interdisciplinary glory” (Estill et al., 2022, para.1).

The source of the initial sample was ADHO’s index of conference presenters. A dataset containing all presenters was organized and sorted into a spreadsheet. An initial pull of the list of the three years of conference presenters gave a sample of 1,277 conference presenters, which was eventually randomized with a final sample of 109 potential interview subjects (n=109). To be eligible to participate in this research, the participant must have met the following criteria:

1. presented at least two papers at an ADHO conference in 2018, 2019 or 2020
2. Confirm ongoing research in DH, or self-identify as a DH scholar

From the 109 potential participants, batches of ten names were selected to be contacted by email. All recruitment was made through the use of publicly available contact information on the internet. If the researcher didn’t have an official university or institutional email address, their name was eliminated (see Appendix A for more detail

on the sampling technique). After the email was sent, if no response was received within 2 to 4 days, a second email was sent as a reminder. Those who did not respond after the reminder were disqualified and the next name on the list was contacted. Out of the total sample of 109, 35 researchers were contacted, and 12 accepted to be interviewed. After the potential participant accepted to be interviewed, a recruitment email was sent (see Appendix B for the participant recruitment email text). The recruitment email stated that the research project was seeking participants with ongoing research in DH, or that self-identify as a DH scholar. This statement acted as a way to screen themselves out if this did not apply (participants were also asked to confirm this at the beginning of each interview). I an email attachment was the letter of consent, and participants were asked to sign before the interview (see Appendix C for the letter of consent). A separate email was then sent as a calendar invitation, containing instructions for connecting to the virtual room in MS Teams.

3.3. Data Analysis

The primary sources of data used for the analysis were semi-structured interviews with DH scholars. Interview data were recorded, transcribed, anonymized, coded and analysed. After transcription was complete, interview data was analysed using TA to identify larger themes in participants' responses.

3.3.1. Thematic Analysis

Using an inductive approach to the data, TA was used as it is one of the most straightforward and flexible ways to analyse qualitative data. The thematic analysis

process was an iterative process – thinking, writing, reflecting, and rewriting. The process meant that I needed to regularly return to the data to inform the coding process. I eventually settled on an organized group of themes, but they continued to evolve until the end of the analysis, even up until completing the discussion section.

The analysis of data in coding themes was based on choices made from a realist perspective, at the semantic level – meaning that the participants were taken at their word. One question asked for the participant's "vision of the future" in DH's data reuse, and responses were then coded at the "interpretive level" (Braun & Clarke, 2006). In a realist perspective, motivations, experience and meaning assume that "language reflects and enables us to articulate meaning and experience" (Braun & Clarke, 2006, p. 85). The themes in this research were identified using Braun and Clarke's concept of prevalence and flexibility to build evidence emerging from the research questions (2006). At times, the themes were inductively measured, and prevalent patterns were coded as themes even if the not all participants made specific reference to these themes in their responses.

Unlike a methodology, which is defined as "a theoretically informed framework for research" TA is more of a method – a "transtheoretical tool or technique" (Braun & Clarke, 2022, p. 3). Braun and Clarke first applied this method to their research in the field of psychology, publishing a description of their recommended process in 2006. Braun and Clarke's approach to TA, which I followed, consists of six steps whereby qualitative data is gathered and the researcher attempts to deeply understand its parts and meaning within the wider context of the data collected. The final step in TA is to

assign codes that describe the data at a micro level, and then probing deeper in order to assign themes that describe the research at a macro level (Braun & Clarke, 2006, p. 87). As a method, TA's aim is to look for and identify patterns in the data without being bound by theoretical frameworks (Braun & Clarke, 2006). It is a common method that helps researchers understand people's "everyday experience of reality, in great detail, in order to gain an understanding of the phenomenon in question" (McLeod, 2001). Unlike grounded theory building, TA allows the researcher to subscribe to theoretical commitments (Braun & Clarke, 2006). The method takes into account that an analysis of qualitative data will feature participants' lived reality and assumptions (Braun & Clarke, 2006). It is a type of analysis useful in qualitative research because, as themes are identified, there are greater storytelling possibilities in weaving narrative detail throughout the research findings (Braun & Clarke, 2006). A thematic analysis of empirical results using an interview methodology often yields richer results than other methods, such as research through surveys or focus groups. Interviews may also allow researchers to "focus on the interviewees' perspective of what is important or relevant, thereby potentially highlighting issues that the interviewer might not have considered" (Young et al., 2018, p. 11).

3.3.2. Coding and Themes

The codes and themes that informed the results and discussion sections of this paper were drawn from interview transcripts. Codes and themes were marked up directly into the interview transcripts that had been imported into a qualitative research

analysis tool called NVivo. To analyse the data, participant interviews were organized under three larger themes – structural and non-structural barriers, and enablers of data reuse. Tables 2, 3 and 4 below show the themes alongside example quotes to illustrate the format.

Table 2

Non-Structural Barriers to Data Reuse

Theme	Example quote
4.2.1. Lack of Time	I wish sometimes we had more time, and we are in under so much pressure to publish. -Dr. Hannah
4.2.2. Accessing the Data	I'm thinking about discarding [a dataset] right now! Often, it's because they're just… they're not well curated. -Dr. Allen
4.2.3. Inconsistent Data Sharing Practices	It's published in papers but mostly, everything is on GitHub. So all the code I use or processes are visible on GitHub. So basically, anyone should be able to reproduce [my datasets]. I mean, it might take some time. But … everything is kind of publicly available and documented… and the code is there for others to reuse. -Dr. Allen
4.2.4. Dirty Data	Like, how do I encode [..] what to me is the same phenomenon in a different data standard? That's really

	complex. There's a lot of issues that make migration tasks really complicated. -Dr. Ken
4.2.5. Permissions to Access Data	[…] I've been kind of like soapboxing, for years now, just trying to convince my colleagues that I work with within digital humanities to frankly, just license their data correctly. -Dr. Jenn
4.2.6. Technical Training Gaps	There needs to be more training and more education on what [data reuse] means, on how to clean data, on how to present data, on how to license data. -Dr. Jenn

Table 3

Structural Barriers to Data Reuse

Theme	Example quote
4.3.1. Labour-intensity	That took an awful lot of manual work. We had like five people looking at that for a weekend, basically. -Dr. Allen
4.3.2. Responsibility of Data Management	I think it's somehow less of a priority for me at the moment, I think, because I think I would see myself more as historian than a data manager... Yeah. Data creator. -Dr. Daniel

4.3.3. Qualitative, Quantitative and Mixed Data	[…] I do not identify as a quantitative or qualitative [researcher] in any discipline. [I'm] not particularly interested in the other categories and labels as much as like, how [I approach] my work and what I think our work is. I do believe, and think that, knowing your methods and foundations matter. Bu that's not how I define what we do. -Dr. Ken

Table 4

Enablers of Data Reuse

Theme	Example quote
4.4.4.1. Collaboration (machine)	So, a lot of these datasets for historians are quite new. And all of these practices are quite new, [so] it's a great opportunity for historians who can also code to make new research on old data sets. -Dr. Connor
4.4.4.2. Collaboration (human)	And sometimes many of the of the projects that we use, and reused the data, [are projects] that we already heard about from other scholars. -Dr. Hannah
4.4.4.3. Collaboration	Sometimes, particularly pedagogically, the existing data sets can really drive what I do, because I need to find datasets that are good [for] teaching. -Dr. Lana

(in the classroom)	
4.4.2. Curated Data Optimizes Data Reuse	When I think of research notes, I'm like … that's data! And that needs to be like, carefully preserved, and you know, sort of organized and timestamped. I don't think a lot of people would think of their research notes as data. -Dr. Jenn
4.4.3. Data Reuse Leads to More Reuse	And then, ideally, though this may rarely happen, somebody else can like pick up that those files, clone them on GitHub, or whatever, and do something else with them, you know, answering questions that the original creators didn't think to ask. -Dr. Boris
4.4.4. Digital Humanists are Autodidacts	I usually write stuff on Python … I'm self-learned completely in programming. I have no formal training in computer science, or computation or programming, which is [the] reason why I'm bit shy to [talk about it]. -Dr. Louie
4.4.5. DH Data Reusers / Vision of the Future	I think DH should [be] leaders and be thinking about data reuse and data sharing and data creation. I think we have a lot of cool data to think about and offer. -Dr. Fiona

3.3.3. Limitations

Although a randomly generated method was used to whittle down a list of potential participants, there are limitations in participant representation based on language, chosen conference, and economic travel privileges. While much of the academic conference and publication world operates in English as *lingua franca,* academic publishing still occurs in non-English languages. Conference proceedings and papers that would otherwise fit within the theme of this paper may not be included in the ADHO archive. Additionally, the participants themselves were internationally represented. This is a good thing for inclusion of different viewpoints across several world regions, but a disadvantage in the sense that there was a slight language barrier between some participants and the interviewer/PI. Another limitation was the decision to use this specific conference to generate the participant sample. While ADHO is an international, well-known conference, it is not the only one to do with DH and therefore potentially excludes scholars whose work was not submitted or accepted. I also want to acknowledge that there are economic barriers to travelling to a distant location for an academic presentation, and this study scope was limited for exclusion of researchers working in cultural and heritage or DH institutions facing financial barriers to access. Finally, my point of view is based on my experience with the Canadian research infrastructure – that lens will have certainly characterized the examples I used and the scope of my analysis.

3.4. Research Ethics

This study uses humans as interview participants, so an ethics submission was submitted for review in May 2021. This research was approved for use from June 2021 to June 2022 by the Office of Research Ethics and Integrity (Ethics File Number S-04-21-6356).

3.4.1. Risks to participants

Participants were asked to read and consent to the terms of the research study through a recruitment and consent letter. These letters laid out the risks, purpose, scope, and intent to publish results (see Appendix B for the participant recruitment email text and Appendix C for the letter of consent). Signed letters of consent will be retained in the medium-term, for a period no longer than two years, in a password-protect folder on the University of Ottawa's OneDrive server. Participants' identities, institutional and geographical affiliations were anonymized and disaggregated, and participants were given aliases. The list of real participants' names, along with a spreadsheet of contact information of participants is stored on the local disk drive of a private computer. This information will be disposed of in the summer of 2023.

CHAPTER 4: RESULTS

This chapter presents the results of the interviews with the 12 participants. The chapter is divided into sections that represent the demographics of participants (4.1.1, 4.1.3), themes as coded in the thematic analysis of transcripts. Since this study adopted an inductive approach, this section features my interpretation before and after quotes to support data collected from participant interviews. The first portion of this chapter presents charts and quantitative findings from the interviews and addresses RQ1. I then address RQ2 (barriers and enablers to data reuse) sorting interview findings into two thematic sections: non-structural barriers and structural barriers. And finally, I end the chapter with my findings on the enablers of data reuse among digital humanists.

4.1. RESULTS OVERVIEW

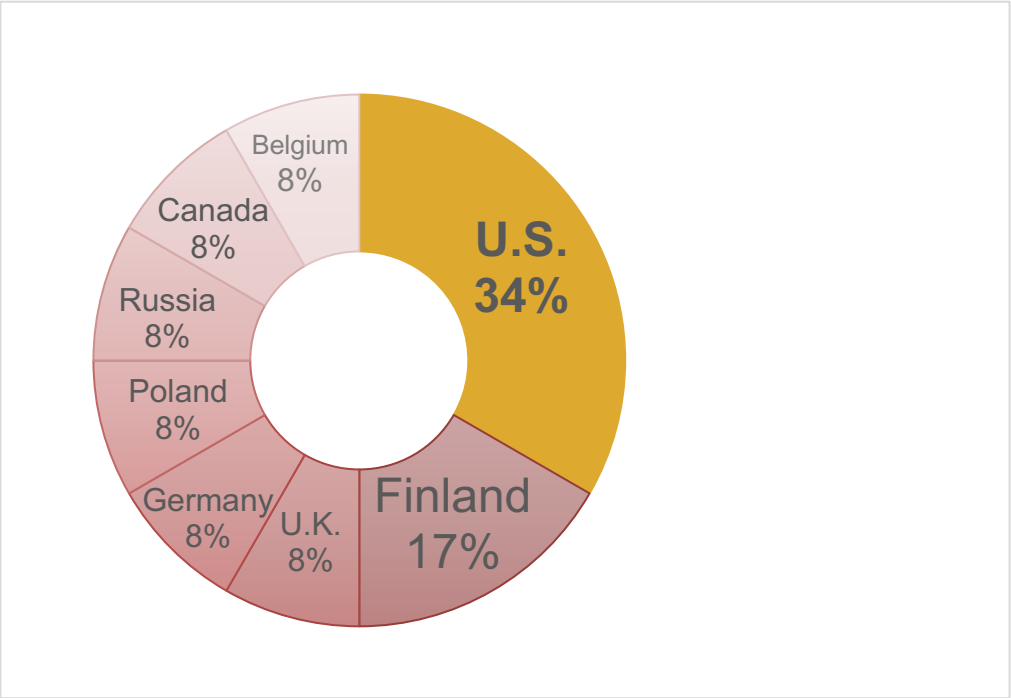
The results show that there are structural and non-structural barriers to data reuse among digital humanists. Those structural barriers include lack of time and resources, accessing the data, inconsistent data sharing, dirty data, permissions to access data and technical training gaps. In the second aspect of RQ2, non-structural barriers include – labour intensity, responsibilities of RDM, and qualitative, quantitative, and mixed data. I structured responses about barriers into the two subtypes – structural and non-structural – while the enabling factors stand alone. Enabling factors include collaboration, curated data, data reuse leading to more reuse, and auto-didacticism.

4.1.1. PARTICIPANT DEMOGRAPHICS

Participants came from eight countries from Asian, North American, and European continents, as shown in Figure 4. The country most represented in the sample was the U.S. (33%, or four out of 12) followed by Finland (17%, or two out of 12). The remaining six were from Belgium, Canada, Russia, Poland, Germany, U.K. Although the sample was randomly generated, the lack of geographical parity in the representation of scholars is one of the limitations of the study (see Limitations, section 3.3.3.).

Figure 4

Participant Demographics, by region



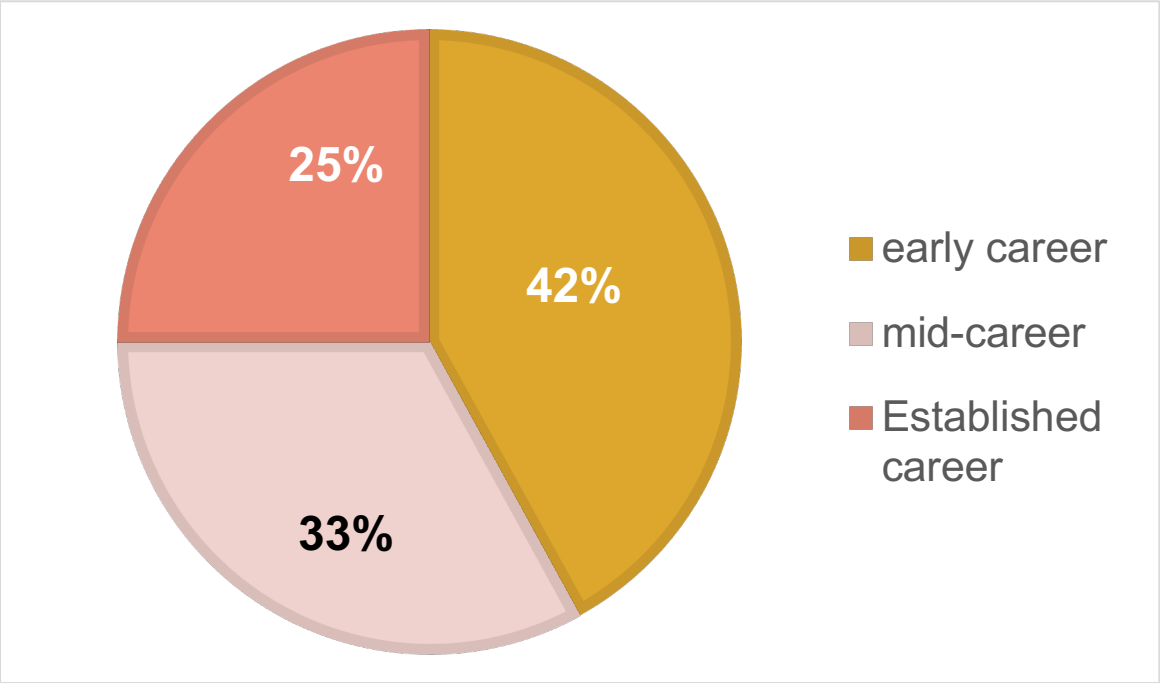
Note. This pie chart shows the breakdown of geographic locations of participants. Four out of 12 were from the U.S., two from Finland, and the remaining 6 participants were from Belgium, Canada, Russia, Poland, Germany, and the U.K., respectively.

Participants from all stages of the academic career path were represented, as shown in Figure 5, the career stages of participants. Early-career represents any scholar completing a PhD or working as a postdoctoral fellow, mid-career represents

someone who has completed their educational path and is working in their field, and established-career represents a scholar with several years of working in their field.

Figure 5

Career Stage of Participants



Note. This pie chart shows the breakdown in participants according to their stage of career. Early career (postdoc or completing PhD), mid-career (working as a scholar for a few years) or established (working as a scholar for several years). The largest percentage slice was 42%, and represents five out of 12, while 33 % represents four out of 12, and 25% represents three out of 12.

4.1.2. What Kind of Data do DH Scholars Use?

All participants were asked at the outset of the interview whether they considered themselves data reusers, and no definition of data reuse was offered to them. Table 5, shown below, lists what types of data participants used or reused in their research. Digital humanists use many different types of data. The scholars in this study said they reused data types such as code, images, text, and none said they had reused audio or video. Two participants said they only work with textual data. Formats of data used/reused by scholars included code, numerical, and data types included forum discussion posts, ancient manuscripts, physical objects such as coins, diaries, and still or moving images.

Table 5

Types of Data Used

Data Type	Number of Scholars
textual	8
code	5
numerical	2
physical objects (e.g., coins, diaries)	2
images	1

4.1.3. Participants' Disciplinary Domains

The participants' disciplinary domains had representation from seven fields: library and information science, linguistics, history, data science, communications and journalism, theology, English (see Table 6). Absent humanities domains were in archeology, area studies (women, black, native, Canadian, etc.), philosophy, performing and visual arts. Most importantly, each scholar referred to themselves either as a digital humanist or DH scholar. Although their other disciplinary backgrounds likely informed their career and research pathways and interests, each scholar had DH as their current central focus. DH superseded all other definitions of their research "selves".

Table 6

Disciplinary Background (after DH)

Discipline	Number of scholars
Library and Information Science	3
Linguistics	2
History	2
Communications/Journalism	2
Theology	1
English	1
Data Science	1

4.2. Structural Barriers to Data Reuse

It is helpful to break down the classes of barriers that present themselves to participants in this study in order to better understand their experience of challenges or barriers in reusing data. Some of the barriers mentioned by participants in their responses can be organized by the overarching categories of structural and non-structural barriers. Structural barriers include institutional, professional, or public structures challenges. These are barriers that the scholar may not be able to resolve by themselves alone, for example accessing resources to decrease time and effort in obtaining desired data, data that is unobtainable for a variety of reasons (access or state of data, inconsistent data sharing practices in academia, permissions to use data, and accessing training to obtain data). Non-structural barriers include a class of challenges that may be more easily resolved by the participants with some adjustments.

4.2.1. Lack of Time and Resources

The most significant barrier to data reuse among digital humanists is the lack of resources. This theme was assigned to interview data whenever participants mentioned sentiments around having to put in time or effort, or feeling time or cost pressures, or not having the necessary resources and support around them (such as help from post-doctoral students, research fellows, research assistants).

Putting in time and effort into managing data is stressful:

> I wish sometimes we had more time, and we are in under so much pressure to publish. -Dr. Hannah

> We had to kind of rework the whole annotation, so it was sounding consistent in the end. So [when I think of the challenges in data reuse], I think this is kind of the biggest challenge. Like, effort of trying to find the right access to a resource. -Dr. Ennis

> It took like, six months just for this particular dataset. -Dr. George

The solution to time and cost barriers to data reuse efforts is not necessarily more time, according to Dr. Connor, because it is difficult to predict how much time a researcher might end up investing, and whether that time will result in locating a suitable dataset.

> You really can't tell before you start doing the research what you actually need from the data…it's going to take time, for people doing similar work as we do but in different fields, to publish their datasets […] the way it is. -Dr Connor

Scholars perceive institutions as more resourced than they are, and believe those larger institutions are better equipped to mine, appraise, select, curate, store and preserve data:

> [What if there are other] organizations, data organizations, like libraries and archives [collecting] similar [datasets]? But it might be that these data need further processing by researchers - it's difficult to predict the needs of researchers for these […] organizations. -Dr. Connor

Cost is an immediately off-putting obstacle when scholars encounter data they are considering reusing:

> [I've noticed there is sometimes] a fee for accessing the data. And this is certainly a huge obstacle in a lot of studies. -Dr. Hannah

Scholars did not have support to maximize their research time – they would rather use the time to focus on understanding the data, as opposed to dealing with technical data access issues:

> Without a postdoc who is sort of dedicated to this issue, people just don't have the time to do that kind of clean. And I think that's a real loss, and we should be investing in that more. But it's not a priority right now for [my university], there's too many other things going on. -Dr. George

Compare wish from Dr. George to the experience of being properly resourced – as Dr. Allen said when he spoke of an anecdote where he wanted to reuse and combine several massive datasets by using the GoodReads.com Application Programming Interface (API). Dr. Allen said he recognized that the success of that project was partly due to him being a tenure-track professor at his institution, and for the graduate staffing support that allowed him to access GoodReads data. He noted that he felt lucky to have had the time of a graduate student to undertake such a big task:

> In that [project], we were creating new data tables [where this] ID matches that ID. […] it involved basically [getting help from] intelligent people looking at web pages and making judgments. -Dr. Allen

Dr. Hannah encapsulates the idea that the challenges are not always about just larger structural barriers, e.g. time, cost, or staffing support. Rather it is all of these parts together that contribute to obstacles in data reuse:

> I think reuse is at the very heart of this age, [and] there are a lot of reasons for this…we don’t have the time, the resources, the manpower, or the space to store all of the datasets that we’re creating. -Dr. Hannah

Scholars noted challenges in locating suitable datasets and believe that institutions, particularly larger ones, are better equipped to handle data management and curation. Costs and fees for data access are seen as a deterrent to data reuse. Data reusers feel they need more hands on deck in order to complete projects that reuse data. Especially in demand are skilled information professionals that understand disciplinary specific data and data management. By comparing the two examples of Dr. George (who at times felt he did not have enough data professionals to support his projects) and Dr. Allen (feeling supported and sharing the labour with graduate students), we can infer that the experience of having graduate or support staff for data compiling, wrangling, cleaning, scraping leads to a more positive experience of reusing data.

4.2.2. ACCESSING THE DATA

Data found on the web can often be unfit for reuse. These reasons vary but the main reason seem to be dirty, uncurated, incomplete, uncontextualized, or unavailable data. Licencing, and gaining permission to use commercial and proprietary presents a big challenge for DH researchers. All interviewees mentioned data access issues as challenges to data reuse. The difficulties in accessing resources to pursue data in academic research was reflected in all scholars working in the academic sector. In terms of a dataset that was downloaded, looked at, and spent some time with, what were some of the factors helping them decide to discard a data set for a research

project? Participants found it wholly frustrating when they could not access data they were considering reusing:

> I'm thinking about discarding [a dataset] right now! Often, it's because they're just… they're not well curated. Or they're not answering the question that I want, or […] it's just the information I thought was possibly there… just isn't there. Or it isn't there in a reliable enough way that I can make use of it. -Dr. Allen

> … we went through the brutal task of encountering the data, it was absolutely unfit for any kind of computational analysis, because it was just scans of very early editions. Very difficult to OCR[…] it was so full of mistakes and of errors in recognition that after trying a few automatic solution for difficult omission of characters, we decided to manually transcribe the texts and verify them [ourselves]. -Dr. Hannah

Scholars find they have to be their own technical support. To solve an issue of accessing a large collection via its API, Dr. George had to teach himself JSON in order to properly "interact" with those API's. In the end, he decided to translate the original XML dataset converted it to a JSON format, and then transform it into a format called JSON Lines:

> I'll be doing the compute [side of a task] and, for example, I'm working on a large project with a colleague that has to do with literary analysis. And our dataset's about 3.4 million records. -Dr. George

Once they searched and discovered data fit for reuse, actually accessing the data they wanted to use was a barrier for a majority of the participants. This participant describes four stops in his search journey, just to access the data to he wanted to mine:

> I would start a search online, I probably wouldn't start up just like a general search, like on Google, I would probably go to GitHub and look to see if there are openly licensed datasets there, that would probably be the first place I start. And other times I've looked at, like Figshare, if it's maybe

> something that's in the social sciences, I might go to ICPSR or, you know, another one of the data repositories [like] Dryad […] -Dr. George

Scholars found they had to troubleshoot technical issues in order to be able to view, assess, or download the data they considered reusing. Once they searched and found a dataset they wished to reuse, the work of accessing the data they wanted to use was a major and oft-repeated barrier experienced by all participants. This is experienced as frustrating and time-consuming for scholars.

4.2.3. Inconsistent Data Sharing Practices

Scholars explained whether they shared data, and if they did, under which circumstances, and where. The multiplicity of storage locations indicates that data is spread throughout digital spaces, makes data discovery difficult. Participants noted that the quality of datasets stored on repositories was sometimes poor and sometimes high quality – but most often it was inconsistent. Datasets were often found to be poorly organized or "dirty", partial, or incomplete. The scholar's own data sharing practices may point to the inconsistencies. Dr. Daniel for example, said he only makes his research data available on his preferred repository, GitHub. Dr. Allen on the other hand, stores on GitHub but will also publish a data paper, linking both to each other in the journal's manuscript and GitHub:

> It's published in papers but mostly, everything is on GitHub. So all the code I use or processes are visible on GitHub. So basically, anyone should be able to reproduce [my datasets]. I mean, it might take some time. But … everything is kind of publicly available and documented… and the code is there for others to reuse. -Dr. Allen

Gaining access to data shared by others shows another aspect of inconsistent data sharing practices. One scholar described gaining access to data from a public institution akin to “jumping through hoops” and said the ease of obtaining the desired data was the most important aspect of his data reuse process:

> So a number of British institutions, which I won't name, because you're recording […] they don't want to share their data. I don't understand that. I don't understand how that works. But what they will do is openly license it, license it but not give you a download link. Right? You have to like jump through a bunch of hoops to try to get their data. So when I look at a digital humanities project, in my like judgy mental filter, the first thing I look for is a button that says download all our data. Right? Or ‘here's the link to our GitHub repository.’ Which means you can download a lot of data. And so I think that's absolutely key. Nothing else matters. -Dr. Connor

GitHub was the most platform that interviewees mentioned most frequently when they were asked about where they shared their data. Some also mentioned university repositories, Google Drive, Slack or personal websites as tools they used to share data.

Interviews with participants seem to point to a shared experience of inconsistent data sharing practices among scholars. This seems true for both the practices of those interviewed and perhaps of those scholars who shared the data that my participants reused or attempted to reuse. Researchers may not necessarily see themselves as part of the research data lifecycle, nor place themselves within it. The scholar’s own data sharing practices suggest a possible reasoning behind the inconsistencies. One participant said he only makes his research data available in GitHub. Another person said they stored data on a personal server, but publishes a data paper that links out to the journal article. All participants stored their data in a large assortment of places

including university repositories, Google Drive, Slack and personal websites. It is interesting to think about all the locations a data reuser would have to search out to find these data, if they were made publicly available. And it is very likely there are even more data storage locations when we think about data researchers are working with at various stages of the research process – working data in active research for example.

4.2.4. DIRTY DATA

Accessing, organizing, and cleaning data was a major barrier and source of frustration among all 12 researchers interviewed. Scholars felt frustrated by the difficulties in finding data they wanted to reuse, their desired research idea and gaining access to the data. Complicating things are human behaviours like their unique, individual heuristic techniques. Each scholar began with an idea they wanted to research and searched an online database for data collections that could answer their research question. Among those places mentioned by participants are GitHub, Kaggle.com, the Internet Archive and the WayBack Machine, academic library databases and repositories, Wikipedia, Wikidata, various archives of parliamentary proceedings, *Rijksmuseum (*Dutch library), the British Library, IMDB.com, Reddit.com, GoodReads.com, bookstores, public discussion forums and the United States Library of Congress. Six of the 12 scholars said they had used APIs, where available, to mine and harvest datasets made available by developers.

Access issues became even more frustrating when the condition of the data was qualified as “dirty”. These datasets were perceived as unusable, at worst, and at best,

undesired. The condition of these were described as disorganized, “uneven”, poorly, or not contextualized, missing code or software, or containing partial or missing data.

> I think that one thing that's true about almost all data sets is that, you know, they almost all have some level of unevenness. And so I think it's, as I'm sure, you know, the data cleaning part of any project is going to be really critically important. Because the way that [the data] may be presented […] doesn't really get into [enough] detail … -Dr. George

> … we had to kind of rework the whole annotation, so it was sounding consistent in the end so that as you asked for challenges, I think this is kind of the biggest challenge like the to find the right access to a resource that you see -Dr. Ennis

> So that's the that's why I think almost 90% of the DH work in general, at least for a historian is just trying to get the data to a standard that can be used. -Dr. Connor

Data that is uneven, or hard to access may even change the research question or approach of the scholar. Dr. Irene had technical issues using Jupyter notebooks she “found online” and said this would mean possibly a change in the direction of her research inquiry.

> I may need to switch a little bit my topic of interest to [match the dataset] and understand better... Right? If the GLAM workbench related to this museum's sphere would be somewhere in Western Europe, I would definitely use it but now it's in Australia and all the data are about Australia and Australian museums. So I need to think [about how] I would like to play with them. -Dr. Irene

Cognisant that scholars will use his data with various questions and approaches, documentation was an important but not central aspect of the dataset.

> If you are going to write a dissertation on letter forms in handwritten Byzantine Greek manuscripts, would you trust our markup on sigma? For your dissertation? Or would you ignore it and read it for yourself? And the answer is no, you would never trust us, because we aren't thinking about the (same) scholarly question. -Dr. Boris

Lack of documentation also led to participants discarding datasets.

> Because the small differences we can manage, but we cannot manage if the data set is just created out of nowhere, and nobody knows how it was created. -Dr. Hannah

Scholars find it difficult and frustrating to locate and gain access to data that match their research ideas. They often relied on various sources that may be uncurated, proprietary, at-cost, or unavailable or undesirable for other reasons. Some used APIs to collect data from sources but this method is not without its challenges. These challenges in data quality and accessibility sometimes prompted scholars to alter their research question. Participants said they often questioned the reliability of datasets lacking comprehensive documentation. This led some to discard datasets, as they could not trust data without proper context and origin information.

4.2.5. PERMISSIONS TO ACCESS DATA

Participants reported being unsure of the copyright permission to reuse data. As they work from different localities in the world, the laws change from place to place. Further complicating is the data provenance versus the researcher's location. Which jurisdictional law should they observe?

It was a challenge for this participant, who wanted to share data while respecting permissions for reuse. Her work uses image files from the U.S.:

> So, the University has a repository, and on GitHub, you can get to it from our [lab's website]. So that's usually what we do if we can share it. Sometimes we can't [share it], and then we just try to explain what we did

> so someone else could replicate it. You can't share TV shows or anything, unless you want the studios coming after you, which is never fun. -Dr. Fiona

There are challenges in obtaining permission to access and reuse certain types of data, especially historical or heritage material, artefacts, and publications.

> [The place where I search for data,] typically, is GitHub. And that's simply because […] if it's on GitHub already, it's easy to fork it, so to speak and to you know, kind of bring it into your environment and work with it. Sometimes, with some of the other datasets that are available, it's … you know… you have to be a little bit more attentive to the fine print that come along with the licenses. -Dr. George

Four out of 12 scholars said they had not continued to pursue access to a dataset because of cost. Often the cost was to buy access from a third-party vendor, who provided a digitization service. Those same scholars mentioned difficulties in accessing material from the British Library (BL). While the BL offers free in person access, there is a cost for digitizing and sending material. And “not everything can be sent by digital file” as Dr. Daniel mentioned, because large corpora would be too expensive because of the cost individual scans of page. Four participants had once attempted to access large datasets from the BL collection and changed their mind because of a cost. Two scholars also faced a geographical barrier – travelling to access the BL collection was too expensive for a research project they had an interest in pursuing. Many participants emphasize issues with proprietary data:

> So, I think parliamentary data is, by nature open data. But for example, now we [are] working with newspapers, [and] the British library doesn't have the capacity to host the data themselves. So, they outsource it to a company for which it is a business model to [sell] access to researchers. And there's a lot of commercial value in the data we [are] working with now, which makes it

> much more difficult to secure. For example, our data is stored like a safe environment, which we actually cannot just […] import and export. [I think they do this] just to make sure there's no data leak, because there's a commercial value in those data, which makes it much difficult to work with. -Dr. Daniel

> [One of the] big problems [in my work] is that the ESTC[4] [English Short Title Catalogue] is proprietary data. So, we have to negotiate with the British Library for the publication and [although] it's going quite well, we want to be in a quite finished state with what we are doing before we start proper negotiations about publishing it. And that's the big problem with many, many datasets, of course. -Dr. Connor

> We work with commercial data, which makes it much more difficult to get access to it. -Dr. Daniel

From Dr. Jenn's perspective, it was relatively easy to obtain permission to use a 16th century English manuscript from the British Library (BL). Her co-worker, an experienced DH scholar, hand transcribed the scholarly edition over a period of time, and another member of the research team worked with the BL to obtain the scanned images:

> I'm not sure what kind of chain of command there was to be honest, I just know that this academic publisher had these scans as kind of like a proprietary collection…? But when we wrote to them, and they just gave [it] to us to use. -Dr. Jenn

She also mentioned that, although it was easy to obtain permission in this research project, she has observed other colleagues in the humanities and in DH struggle to obtain permission to reuse data. And even though these scholars struggle in their pursuit of permissions to use a desired dataset, she opined that those same

[4] The ESTC is a database listing 480,000+ publications from the 15th to the 19th century.

scholars were not necessarily any better at sharing or licencing their own collected data later on:

> I've been kind of like soapboxing, about that for years now, trying to convince my colleagues that I work with, within the digital humanities to frankly, just license their data correctly. Which doesn't happen, because … .it doesn't happen for a lot of reasons. But I think largely it doesn't happen because people are kind of working on their own projects, so they're like, "why do I need to think about licensing data? This is my data for my project, I'm just working on it." It's not that they're like hoarding it or anything, they just don't necessarily, they're not necessarily thinking in the reuse paradigm. And so, I do a lot of talking with folks about "No, it is actually really valuable to have license your data… and these are the reasons why." -Dr. Jenn

Dr. Irene said her preference was first to look for datasets available for reuse, "I can download something published by my colleagues under a free license, under open […] licensing". However, if the right dataset did not present itself in her search, she would then collect her own raw data.

Obtaining permission to access and reuse certain types of data – such as historical data, or heritage material and physical artifacts – was challenging. Scholars often encountered uncertainty regarding copyright permissions for data reuse, posing a significant barrier and leading to the scholar's frustration. Participants were often unsure which laws to follow, considering their geographical location and the jurisdiction of the data. Cost was another barrier to accessing data, with four participants noting that they had abandoned the pursuit of a dataset due to financial constraints, often related to proprietary data with closed or restricted access. Overall, researchers preferred to reuse existing datasets with open licenses.

4.2.6. Technical Training Gaps

Access to training and lack of resources like time or money were major challenges and often mentioned in the same sentence – technical training needs beget resources in time and money. Participants often wrote their own code or create their own tools to exploit, view or analyse the data. The theme of being self-taught is coded as both a challenge and an opportunity for scholars interviewed for this study – its positive attributes are explored later, in the section on digital humanists as autodidacts in 4.4.4.

The majority of participants wanted technical training that would help them manage their data more effectively:

> We need to guide especially those who work in the humanities. Usually, the skills [needed] are very, very basic to play with data. But not many historians […] can build their own dataset at the [appropriate] level and properly describe this dataset. -Dr. Irene

> There needs to be more training and more education on what [data reuse] means, on how to clean data, on how to present data, on how to license data. -Dr. Jenn

There is a need for guidance in reusing and managing data in the humanities. Additionally, there is a need for more comprehensive training and education on various aspects of data reuse, including data cleaning, presentation, and data licensing.

4.3. Non-Structural Barriers to Data Reuse

These barriers are organized around human factors such as attitudes, motivations, process, habits. Non-structural barriers to data reuse appear to be tied to sentiment, attitudes, and perceptions such as labour intensity, responsibility of data

management, inconsistent data sharing practices, difficulties in accessing the data and diverse and wide-ranging data types.

4.3.1. LABOUR-INTENSITY

Reusing data is time-consuming. Some oft-cited benefits of data reuse are a perception that the time spent not collecting data will save time and money, yet research shows this may be untrue (Pronk, 2019). Scholars interviewed for this thesis project felt the tasks of searching and discovering, cleaning, contextualizing, and organizing data for reuse was labour intensive. Participants stated that there was much work to be done at some stages of the research lifecycle, for example the planning and collecting stages or the or reusing / creating / sharing stages:

Tasks that were noted as difficult in creating the dataset included cleaning, migrating, scraping:

> As the standards evolve, you have to migrate all your data if legacy data is outdated, or no one's going to use it anymore. And that is a major technical issue that touches so many other issues, like funding. Like, where do we get money for proofreading [datasets]? And how do we make sure that our data is still the same? -Dr. Ken

> Gathering the corpora is the most time-consuming part of every study that we do… gathering the data on our own for each project would be simply impossible. -Dr. Hannah

> Researchers have a lot on their plate… and this is another thing to add, right? -Dr. Jenn

Labour intensity is linked to the lack of resources theme identified in the section under structural barriers (4.3.2). The time, cost or staffing support can be identified by the scholar, thereby presenting an opportunity for them to request support (such as

posting a position for a research assistant or funding for postdocs). If the funding is not granted or an assistant is not found, the barrier then becomes a structural one. As Dr. Jenn remarks, if DH scholars are going to reuse data reliably and consistently, "it has to be really easy. But it also has to be linked to [things like] like funding". The idea of funding is interesting. It suggests that the labour intensity barrier of reusing data could be lessened with structural changes at the funding policy and institutional level. Lessening the pressure to obtain funding would help reduce frustrations among scholars since securing limited research funds is a competitive and time-consuming process.

4.3.2. Responsibility of Data Management

Tasks that were noted as difficult that also fell within the research data lifecycle, or RDM, and there appears to be various attitudes about roles – who should take on the responsibility of the later stages of the research lifecycle – such as sharing and preserving data for the long term. Each scholar had a different perception. When asked about their RDMRDM practices, every participant except for one, Dr. Hannah, said they did not believe that data management was part of their role as a scholar:

> The challenge, though, is that we all do [RDM] in different ways. So there might be people who will prefer to use GitHub like myself, or who prefer to use older […] or other platforms for sharing. But [deciding on a repository] across all of these platforms can be time consuming, especially since not all of them are really fit for browsing. -Dr. Hannah

Participants viewed themselves as having either a hands-on or hands-off role in the data management of their research projects. Others felt they were already experts at managing their data:

> I am sort of an expert in that. So, I would probably be more likely to create knowledge [tools or tutorials] about [RDM]. -Dr. Allan

> I mean, at this point, we're […] pretty good at this stuff, right? Like, you know, I've been doing it for a long time! The thing that would be most valuable would be somebody who understands what [the funder] wants to see in our data management plan. It's very cynical, right? Like, I know how to manage our data. -Dr. Boris

> And I think, you know, a lot of the stuff really needs to be done at the practical level, just so that you work through all those data, cleaning issues and version issues yourself. -Dr. Hannah

> [Our library has been] really suggesting that [RDM] should start at the beginning of your grant writing cycle. […] That's ideal. [They recently] presented that famous sort of data curation lifecycle […] it shouldn't be an afterthought. -Dr. George

One scholar thought RDM was an important part of their role as DH scholars:

> Part of my role is being the data manager. -Dr. Connor

One participant described himself as more of a data creator, as opposed to a data manager:

For Dr. Irene, attitudes towards data management are regionally specific:

> Data management is not big in [my country]. I bet nobody cares. – Dr. Irene

One scholar felt that it was important for scholars to see themselves as stewards of their own data, and that RDM should be systematically taught to new generations of doctoral students in her field:

> [Here's] the problem that I see, [and it's] very painful to me as a researcher [to say this]. While a lot of young researchers take great care [in] maintaining their online persona, providing links to their research, and so on, when it comes to research data management, I think this is something that's not being taught enough to the PhD students. And it's a very important topic. It's crucial for our comfort basically, [or at least it is] when it comes to my own work. -Dr. Hannah

Two participants believe DH scholars should include the role of data steward in their work, and work on developing good RDM practices:

> We should really be focusing on data first rather than interfaces in digital humanities. And so I think that, too often what happens is people get really fixated on the appearance of what the [work] should look like. And then they don't spend nearly enough time actually curating and making their data available. So this is sort of a kind of plea to, to focus on getting the data in order before you work on the interface. -Dr. George

> It's something that needs to be kind of baked into project design and thought of throughout [the project. We ought to be] trying to encourage people to think of themselves as, you know, data stewards, and less as just like, hey, I made this thing someone else take care of it, you know? -Dr. Jenn

These findings about RDM responsibilities suggest that data stewardship roles are not clear for scholars. Clarifying and defining these roles would allow scholars to spend more time focusing on research and less on data management. As one participant (Dr. Jenn) astutely notes, RDM must be planned in advance of undertaking new research projects. So while RDM support would be great to have, scholars should also learn and apply those skills early on in each project cycle, beginning with Data Management Plans, to "bake in" RDM into the design of the research. Dr. Irene points out that RDM is not popular in her country. This suggests a deeply ingrained culture of academic research, one that goes beyond geographical borders. And so, while the

responsibility of RDM might be a non-structural barrier that could have simple remedy, it is also a structural barrier that will need to be removed with a massive cultural shift in science in practiced.

4.3.3. QUALITATIVE, QUANTITATIVE AND MIXED DATA

Participants spoke of a wide range of types of data they worked with and struggling with the sheer amount and diversity of formats available in the research data landscape. Far from just being described as either quantitative or qualitative, all participants used a mix of both. Their data were often a mix of textual and numerical data.

> So, I think having our disciplinary sensitivity is really important. I don't necessarily think we need to break down into sub fields and say, like, okay, DH, can only have *this* kind of research, data management and history can only be *this* sort of research. -Dr. Jenn

> [There are] quite a lot of cool approaches on how you could on how you could categorize data. Dr. Ken

In searching for data, participants designed broad research questions first and went looking for suitable data to answer their questions secondly.

> […] I do not identify as a quantitative or qualitative [researcher] in any discipline. [I'm] not particularly interested in the other categories and labels as much as like, how [I approach] my work and what I think our work is. I do believe, and think that, knowing your methods and foundations matter. But that's not how I define what we do. -Dr. Ken

> I've probably used almost every form of data, I tend to be focussed a little bit more on the quantitative side, and at least for myself. I don't do qualitative data analysis […]. Dr. George

> I think that, you know, for quantitative data, it comes in lots of different

> forms. We've got a book coming out [that includes] metropolitan data, and other map data. One dataset we're using [is about] TV sitcoms, another is on movie posters. And another is [about] discovery of scale for [cultural] institutions. So, we're kind of all over the place and all of it is on a broader scale of the work. -Dr. Fiona

One participant said that DH is such a unique field because of its ease in going between qualitative and quantitative data, and vice versa:

> DH – by its very nature of looking at humanities data over centuries of humanities, culture, and civilization – offers researchers the chance to turn numbers into a narrative, or in the inverse, to quantify qualitative data. -Dr. Connor

DH is often at a crossroads of qualitative and quantitative data. Part of the problem in normalising data sharing and data reuse as a regular part of research tasks could be linked to an identity crisis within the humanities. The humanities discipline is sometimes perceived as not being a “real science” and as Dr. Jenn notes: “humanities kind of seems to be under perpetual threat”. The most fascinating part of this finding is that, by designing broad research questions first and looking for data second, scholars change or adapt their research questions according to data availability.

4.4. Enablers of Data Reuse

In order to assess what kinds of enabling factors lead to data reuse, interview data were analysed as a whole as opposed to one or two specific questions. Just as the barriers to data reuse interviews were analysed for sentiment, attitudes and motivations, this section is an overview of their responses as interpreted by the interviewer. The aim was to gain a sense of their experience of data reuse where it was working well. Some

participants naturally offered these glimpses into their positive experiences, and others had to be prompted when the information did not come up in conversation.

The main themes in data reuse opportunities were collaboration, curation for data optimization, reuse leading to reuse, the ethos of digital humanists as self-taught, and their own enabling practices of reuse.

This section deals with the question posed in RQ2 and is ordered in five sections of themes within enabling factors to DH scholars reuse of data. The collaboration theme include responses mentioning connections, linkages, communications with other people in regard to the participants' work. Technology, Tools, and Machines is about meaningful engagement with data science tools and Artificial Intelligence, Optical Character Recognition (OCR), the Text Encoding Initiative (TEI), XML, and other tools.

4.4.1. COLLABORATION

Participants collaborated with librarians, their students, and other researchers – both those they knew and those whose data they came across in a search. They also "collaborated" with and made use of technology tools as needed. The literature suggests that both students (Tewell et al., 2017) and scholars (Grossetta-Nardini et al., 2019) do some form of consultation or collaboration with librarians at many stages of research workflows. Participants were asked if and how they collaborated with librarians, but the interview questions did not specify in what capacity (for example, providing services to researchers such as helping with literature or data searches, or conducting scoping or systematic reviews). Scholars said that instructions on how to

gain access to data collections was the main form of collaboration with librarians. Other did not mention working with librarians. It may be a that researchers don't know what services are offered by librarians – for example, helping with research, or with research practices like RDM. Two of the scholars interviewed also had experience in librarianship, and seemed to want to know how to better promote the role of the librarians to the scholar:

> [RDM] is not something that librarians are just going to be able to swoop in at the very end and like, you know, put a little stamp on [the project]. -Dr. Jenn

> [The way scholars seem to learn about RDM workflows] might be you know, being like a PhD student in a lab and having other students sort of showing you, here's how we manage our data, here's where we share it, you know, and, and you learn the tools that way. And that's both good and bad. I mean, it means that everybody's not necessarily following the best practices all the time, because, you know, some are a little bit more attentive to what those practices are, and others are not. And so I think there's an opportunity for the library to kind of to, to listen in on what's going on in those different areas, and then try to, like transmit best practices across disciplines. -Dr. George.

One scholar said he had a positive experience with a librarian in trying to access resources at the Dutch National Library:

> […] you can actually work with your data in collaboration with the research engineers to get feedback for how scholars can […] work with a collection basically. And, for example, in my case, I was interested in researching gender bias in newspapers. So then they would basically give me access to all the newspapers. […] So it was quite an open, quite a good relationship that I discovered, to make sure you get the data. -Dr. Daniel

Dr. Connor said he actively sought out help to access a very large data resource from librarians:

> The second-best scenario where I might ask for help, is [from] the library in some way. And if [the dataset that I want] isn't available as a dump, then I would need some services from the library to prepare a data dump for me or instructions on how to access the library API. -Dr. Connor

Scholars said that collaboration was often in the form of sharing resources and discovery tools to gain access to data sources. They often learned about a data tool or source from other scholars they work with, or that they knew from social media, conferences, or published articles.

> […] sometimes many of the projects [where we reused data, we had] already heard about from other scholars. When it comes to looking for the data for completely new projects that we embrace, like, very often I will rely on the knowledge of people I cooperate with. So if I'm working in the language that I don't speak or that I don't have much experience with, I will just work with a co-author who [is] more versed in the specifics of [the] collections that can provide us with the data. -Dr. Hannah

> Do you know [American scholar]? She's a really engaged person on Twitter, and she sent me information… and links to data. So if I want to, I can find some data, and some [of it] is very beautiful. -Dr. Irene

Participants in the study seemed eager and willing to collaborate with students and fellow researchers, and some of the participants mentioned collaborating with librarians. The collaboration experience took many forms (email, social media, in person) and had varying functional purposes (e.g. sharing discovery tools, or data wrangling, mining, or cleaning tools). For participants who had worked with librarians, these collaborative experiences often involved learning about new resources and tools for discovery and exploitation of data. It is also interesting to see how scholars discovered new data sources. Discovery came from many collaboration routes, including through colleagues, social media, conferences, and published articles.

4.4.2. Data Reuse in the Classroom

Almost half the participants (5 out of 12) had teaching duties as part of their professional duties. participants said they search for datasets not only for their own research projects, but seven out of 12 scholars said they needed to find data to reuse for their students as a teaching aid. This subset of teacher-participants all said they need regular access to a collection of reusable datasets in their teaching. Their reasons for doing this varied, but generally they wanted to help their students save time by culling and sharing a list of repositories with large datasets.

> I teach [several] classes, so a lot of times, I'm sort of coaching students in how to find datasets, I really don't want them collecting things themselves, because that would take too much time. -Dr. Allen

The participants with teaching duties not only reuse data themselves but are demonstrably invested in sharing their enthusiasm for data with their students.

> We do draw on data that's been produced by third parties quite a bit in teaching. -Dr. George

One participant said she was particularly interested in finding “fun” datasets for her students to “play with”. Dr. Lana said that she prefers the save herself some time by looking for datasets and repositories that are both good for teaching purposes and relevant to her current research projects. The datasets she looks for must therefore serve double duty, and ultimately affects selecting or discarding datasets:

> Sometimes, particularly pedagogically, the existing data sets can really drive what I do, because I need to find datasets that are good [for] teaching. -Dr. Lana

She goes on to describe barriers in finding datasets suitable for teaching her students:

> Datasets I need for [teaching] are hard to find because of the proprietary nature of them. And they're more 21st century data ... they have old companies do their data's work. Some [data producers] have done very good job at making the data open and available. So that's kind of why I end up sometimes on Kaggle or like IMDB or those kind of sites for grabbing their data. That's because this search of data is broad enough [for] most students [to] understand. -Dr. Fiona

Data sharers do not make their data easily or freely available, which leads to less use by emerging scholars and students. Participants experienced difficulties in accessing proprietary data, not only in a researcher setting, but in a pedagogical setting. As pedagogues, some scholars wanted to make the experience of working with data a positive, even enjoyable one for their students. This was deemed often impossible because of the costs associated with proprietary data. As DH scholars are enthused reusers of data, reuse is enabled by their positive experience of working with datasets, but dampened when data access to particularly interesting but proprietary is limited or closed.

4.4.3. TECHNOLOGY, TOOLS, AND MACHINES

Almost half of the participants have technical expertise (5 out of 12) and three have computer science (CS) backgrounds.

> So, I've been involved with digital humanities for longer than I care to admit. At the same time, intellectually, I'm sort of an Information Science, AI machine learning person, fundamentally, [but] I sort of describe myself as: 'I speak humanist, but I have a strong CS accent'. -Dr. Allen

Certain tools enable humans to work with machines and enable humans to create metadata and new machine-readable data from data collected by others. One scholar talked at length about leaning on the collaborative community of Text Encoding Initiative (TEI) enthusiasts. TEI is collaborative system for digitally describing texts in the humanities using Extensible Markup Language (XML). It links a consortium of users who together share the responsibilities of developing and maintaining the standard for the representation of texts to make it machine-readable. Dr. Ken, who researches historical handwriting, believes that DH is constantly being improved because of its collaborations amongst digital humanists, and that this collaboration optimizes digital data so that it is more reusable:

> If one looks at how [my] projects are related to one another, it's also produced by the same people that informed the TEI. You know, [there were] all these people developing annotation schemas back [in the 1980s]. Some editions in [my home country], for example, […] they're producing print editions, digital scholarly editions, and […] at the same time are members of the TEI community. Some of them are really active members, shaping how the TEI looks. -Dr. Ken

One respondent spoke of the collaborative across generations of people doing DH since the 1980s, with three out of 12 participants citing TEI. The nature of the work of digital humanists is such that they build on the work of collaborative tool like TEI:

> How do we annotate, cross out deletions, in texts, meaning textual development or produce another standard? […] Over the years, I noticed people working together and then saying, hey, we want to build this one thing and embed data. And so these people are all connected. It's, you know, it's not like the people in the 1980s worked completely disconnected from us today. You know, they built the groundwork of what we're doing today. -Dr. Ken

Another participant said that he found mark-up and annotative tools helpful to him as young researcher when he was working with a Greek language corpus but did not speak it (Dr. Boris). Annotations for language, and other metainformation about texts may similarly help other researchers who have little knowledge of the language of a foreign research corpus. Knowing how to code or program can be a new text mining opportunity for digital humanists working with historical documents:

> So, a lot of these datasets for historians are quite new. And all of these practices are quite new, [so] it's a great opportunity for historians who can also code to make new research on old data sets. -Dr. Connor

Certain tools provide the entry point between a researcher and the tool designer. When she couldn't find the museum data she wanted from the Internet Archive, Dr. Irene learned of the "GLAM Workbench". This tool, created by an Australian researcher, helps non-technical users harvest large collections from the GLAM sector (galleries, libraries, archives, and museums). Users can learn or share tutorials on how to scrape from cultural heritage collections. Dr. Irene is a self-described non-technical scholar, and though she wishes to learn to use machine learning tools and programming language to better access, use and exploit data in her research and for teaching, for now she relies on the collaboration and collegiality of social technical tools like the GLAM Workbench:

> I'm on the crossroad. So, I would like to reuse data. But I need very particular […] datasets in general because I need something related to web museums. […] I know about the GLAM workbench in Australia, [and I] actually even know the person who made the GLAM workbench. I'd love to use this data. -Dr. Irene

As a collaborative activity, reuse practices are also an opportunity to practice making reproducible science:

> I think that [data sharing and reuse] kind of moves in a much more, you know, sort of complete sense of data. Data reuse [flows] in the direction of data and reproducibility. [Currently], there's an effort in [my] library to sort of collaborate across the whole university with various folks that are really invested in reproducible data. -Dr. George

There is clearly so much opportunity in collaborative mark-up and annotative tools that allow scholars from different geographic locations, languages and even generations to communicate and work together. What new tools could be used to capitalize on DH scholars' use of documentation and linked data?

4.4.4. CURATED DATA OPTIMIZES DATA REUSE

When data was curated, it contributed to a smoother experience for the researcher. Data was more likely to be selected because it appeared to have been curated (metadata assignment, documentation, file contextualization, etc.). This made the data easier to find and to understand. When asked where he finds data to reuse, Dr. Allen said he looks for well-curated data collections, on a user-friendly discovery interface:

> [I tend to find datasets I want to use on] Kaggle. They're sort of very curated. There are other cases, like, I think Harvard has like a Dataverse or something like that. [It's a] repository that I've used. More often, it'll be something where someone has created a dataset and put it somewhere. -Dr. Allen

> I think a lot of the actual work is really about finding, curating, and reusing datasets. Figuring out like, what, are the affordances of this data set? That's kind of how I've started thinking about it. Like, what can I do with it? What are the handles on it? How heavy is it? -Dr. Allen

Data curation includes contextualizing the information in a dataset so that a potential reuser can understand and make sense of the data. Dr. Ennis emphasized the need to see a guiding document such as a data dictionary, or a ReadMe, appended to the dataset files, for example, so that he may better interpret and contextualize the annotations made the previous researcher:

> One challenge is to properly interpret what you see. [For example] say I'm not capable of Sanskrit or something. So, I kind of have to rely on what the researchers [..] did. And [in this case] what they did was they annotated things. So, for example, the dictionaries need to have categories and I have to understand the structure of the thing. So, I have to rely on not [only] consistent annotation, but also interpretive annotation. -Dr. Ennis

> When I think of research notes, I'm like that's data! And that needs to be like, carefully preserved, and you know, sort of organized and timestamped. I don't think a lot of people would think of their research notes as data. -Dr. Jenn

Professional data stewards or curation specialists would be welcome guidance for researchers such as with this participant:

> We need to guide [people], especially those who work in the humanities. Not so many historians can build their own datasets at the […] proper level and properly describe this dataset. So, it [needs to] be managed in some way, or it can be done by professionals. -Dr. Irene

Curated and well-stewarded data are more likely to be reused by scholars because of the elements that made them findable in the first place – metadata, documentation, and information on the provenance. This is interesting because it is another example of an instance where these practices need to be in place at the time the original data collector shares the data, showing that data sharing and data reuse are inexorably linked.

4.4.5. Data Reuse Leads to More Reuse and Richer Data

Some scholars said that they are open to sharing their newly enriched datasets, but sometimes they were not sure of the permissions and copyrights. Reusers tend to remix data – patching together datasets to make a new or larger dataset. This patchworking of data *enriches* research by virtue of the fact that academics continuously contribute to the corpora of knowledge in their area of expertise. When participants had a positive experience in reusing data, such as the ability to add to or enrich the data, the more the participant reused. That could be caused either because of the scholar's experience (the process led to good or interesting findings), or perhaps because this method becomes their preferred way to conduct research. Participants were excited by the potential for harvested data to be reused in more than one way in research, simply by asking a different research question. They were happy to combine datasets to create new, enriched datasets, answering new questions of the data:

> And then, ideally, though this may rarely happen, somebody else can like pick up that those files, clone them on GitHub, or whatever, and do something else with them, you know, answering questions that the original creators didn't think to ask. -Dr. Boris

> Our [English Short Titles Catalogue] datasets are so rich, that it isn't empty in one historical question – you can milk it for more and more historical research. So [it's] definitely being reused. And we've already published historical stuff based on that. So, it's being used in research and then used more, and it can also be just pivoted to another question even without further developing the dataset. -Dr. Connor

Dr. Hannah studied an anonymous Mexican play attributed to a famous writer but lacked Mexican texts for comparative purposes, so she searched for resources and

found a Spanish Baroque collection and digitized libraries, but only part of the collection fit their criteria. Her and the research team had to then find a way to augment the dataset, eventually finding that their library held digitized collections of relevant data. In finding additional data to augment the existing data about her subject, Dr. Hannah created a new dataset. If she were to share this larger, enriched dataset about language in the Spanish baroque age, this would be a contribution to and enrichment of the knowledge infrastructure.

> I think when we extend the corpora, for example in the context of text. If we are sharing it, and it includes more information, or more data points then it becomes […] this new version. -Dr. Hannah

The question of who owns the intellectual property and how to manage copyright permissions of data was not clear in four out of 12 participants, who said they believed they were the author of a dataset they had reused and enriched:

> I would say yeah, correct [I am the author of the new dataset]. And in most cases, we do exactly that [we are] using resources to enrich other resources. So that's kind of the most common way to reuse data in our context. Like for example, in digital editions of existing texts [data reusers] enrich the texts and use other resources to enrich it. [They're] creating a new, more enriched metadata data sets out of [my data]. -Dr. Ennis

> [When I'm going through the datasets I want to reuse], what I've been doing is trying to create more than datasets out of that publisher and book data that you can get out of [it]. […]. So, I've been creating a new, more enriched metadata dataset out of [the original data]. -Dr. Connor

Certain participants hesitated to "claim" IP over an enriched or augmented reused dataset. One scholar used the word "extend" to describe enrichment she performed on the original dataset. Dr. Hannah said she would potentially claim a

contributor's role or some other form of authorship if she felt she had sufficiently extended and added to the corpora. For Dr. Hannah and others, the *extent* of the work it took to extend, enrich, or augment the dataset was key to whether or not they would share the enriched dataset and add their name as contributors to the dataset:

> I guess I would push back on the definition of [authorship]. Like how much am I reusing, versus sort of collecting something that was collected in some way. Like, a lot of times, we'll hit an API a whole bunch. So that's technically a dataset that existed but we're sort of creating a new dataset. -Dr. Allen

Dr. Irene ended up building "a mixture of data" that combined numerical and textual data, some reused, some combined.

> [The hard part about] this model of data was more about trying to find the appropriate features, to describe them in the same, you know, at the same level. And we tried to organize this information [in a coherent way] by using a TEI file, to explain the main semantics of the file. -Dr. Irene

As Dr. Hannah points out, DH can be leaders in exploring the idea of data authorship:

> […] this is something that's being discussed in the digital humanities quite a bit as well: [when it comes to sharing the data], who is the owner of the data? -Dr. Hannah

The scholars' knowledge and opinion of copyright and intellectual property concerning reused data was tenuous. Some participants were unsure when and if to "claim" rights over datasets. Four out of the 12 were new scholars, many working as postdocs or completing PhDs. Being able to add data co-authorship to their accomplishment could potentially be a motivating factor for scholars to share their data.

4.4.6. Digital Humanists are Autodidacts

Digital humanists train themselves. Working professionals often upskill to remain current in changes in their field with new training – both to learn and to remain up to date in the training. The participants in this study said they actively sought out to learn on their own how to use a computational tool to would help them access a dataset they were coveting. These included learning to using new research software or programming their own, learning to code or a programming language like Python, creating programs to clean, wrangle, exploit, access or view the dataset, as well as annotation and encoding tools like the Text Encoding Initiative (which was developed by the DH community in the 1980s, and has been maintained through TEI enthusiasts in DH like Julia Flanders and others).

Reusers in DH are often self-taught:

> I usually write stuff on Python … I'm self-learned completely in programming. I have no formal training in computer science, or computation or programming, which is [the] reason why I'm bit shy to [talk about it]. -Dr. Louie

For one scholar, learning advanced computational skills was not yet a reality. When asked if she was comfortable using different programming language like, Python or R:

> No, no… that's why we have to improve our digital literacy. Our professional literacy. But I'm in the very beginning of this process, even being a

> professor in digital humanities, it's a shame, but I have forgotten [how to do these] things. -Dr. Irene

Yet Dr. Irene later says that she worked on a highly technical project developing a TEI scheme, documenting the main semantics of the project.

> It came out of librarianship, I have to say, you know, honestly, […] there's just so much to learn. You just can never really catch up; you can just barely tread water. And, and it's really driven by the needs of researchers, you know. -Dr. George

> Everything that, you know, I've been talking about, it's all been connected to the actual practical research needs. And I think that part of the job of a kind of contemporary librarian, especially if you're in the digital arena, is just to try to stay current as best you can. And, you know, the way that you do that is by, you know, maintaining context with researchers across lots of different disciplines. -Dr. George

Training needs to be asynchronous, self-paced and easily accessible:

> An asynchronous resource would be more of preference, I don't really have the time to follow […] live events. I sometimes listen to online lectures, but only if they are of much interest. But I do online courses definitely. -Dr. Irene

> I would prefer to [learn through] some sort of guideline with descriptions, like ‘how can we download databases to work with APIs’ [for example] -Dr. Irene

Digital humanists are often self-taught, relying on self-directed learning to acquire the skills they need to stay current in their field. They actively seek out resources to learn how to use computational tools, such as programming languages like Python or specialized tools like TEI. Many of them actually lacked formal training in computer science or programming but were motivated to acquire these skills to advance their research or careers. Participant interviews suggest that technical and RDM training for digital humanists will have a better chance at success if it is presented asynchronously,

is self-paced, and is flexible enough to accommodate the busy schedules of working researchers.

4.4.7. DH Scholars are Data Reusers

DH scholars themselves are enablers of data reuse and want to not only themself reuse their own data but want their project collaborators and close colleagues to reuse their data in the near and far future. All participants voluntarily revealed, at various points in the interview, that open science was important to them. This value seemed important to them in two important ways. First, in the sense that they had personally experienced barriers in accessing data, and two, in the sense that *they themselves* intended to reuse their own data:

> The whole point of this [project] was aimed at data reuse. And largely just internal to our project. We were going to edit texts from manuscripts, we were going to edit texts from papyrus. We were going to edit texts from other editions at the outset, not knowing what we were going to find, and therefore not knowing how we were going to want to mix and match this stuff. So even within our project, it was a data reuse challenge. Like, we're gonna do a hell of a lot of work. And in 10, or 15, or 20 years, we're gonna want to analyse this data. So, we have to be able to reuse it ourselves. -Dr. Boris

Digital humanities (DH) scholars are enthusiastic proponents of data reuse, and this commitment is exemplified by Dr. Boris, who emphasized the importance of data reuse within his project and the necessity of being able to analyze the data they create in the long term. Initially focused on editing texts from various sources like manuscripts and papyrus, Dr. Boris said that his lab's multi-year project eventually centred on optimizing data for future data reuse. The research team anticipated the need to reuse

and repurpose data early on and recognized the importance of addressing data reuse challenges within their project. The motivation was to ensure that the extensive work they were doing could be analyzed and reused effectively by them and other researchers in the coming years. Dr. Boris said he used various strategies to optimize data for reuse, including documenting workflows, creating metadata records, and utilizing mark-up and annotation within the data itself. Clearly, DH scholars not only desire to reuse data themselves but think sustainably in terms of data legacy and stewardship. Modelling and encouraging collaborators and colleagues to adopt good data reuse practices is also a great example of how culture change can happen at a granular, local level.

4.4.8. Data Reusers in DH have a Vision of the Future of Reuse

Participants were asked an open-ended answer to close out the interview — did they have a vision of the future for digital humanists working with reused data? This improvised question yielded rich, speculative responses.

> There's kind of this disconnect between open access publishing, and open data practices. [DH] isn't thinking about open scholarship as an ecosystem […] It's all of the opens. -Dr. Jenn
>
> We should really be focusing on data first – rather than interfaces – in digital humanities. -Dr. George
>
> I think that reuse is at the very heart of [this] age. […] We don't have the time, the resources, the manpower, or the space to store all of the datasets that we're creating. -Dr. Hannah

> I think DH [researchers] should [be] leaders and be thinking about data reuse and data sharing and data creation. I think we have a lot of cool data to think about and offer. -Dr. Fiona

DH data reusers are on the vanguard and have a unique perspective on the future of data reuse. They see a disconnect between open access publishing and open data practices and emphasize the need to consider open scholarship as a holistic ecosystem. Some suggest that the focus should shift from data tools and interfaces to prioritizing data itself. There's a recognition that in the current age, data reuse is essential due to limitations in storage capacity and high-performance computer resources. Digital humanists could take a leadership role in promoting data reuse, sharing, and creation as they possess valuable context and experience.

CHAPTER 5: DISCUSSION

This research project set out to answer two main exploratory questions: How do digital humanists use, find and reuse data (RQ1), and their barriers and enablers regarding data reuse (RQ2). I used interviews to collect data from a randomly sampled group of DH scholars, and asked them a series of questions about the challenges and opportunities these scholars might face. The study found that the present data reuse infrastructure is made up of both challenges (structural and non-structural) and opportunities (enablers). The discussion that follows includes reflections and six recommendations for improving the infrastructure and are centred on education, data discovery, and the culture of the research infrastructure.

5.1. KEY THEMES AND RECOMMENDATIONS

The following section discusses the findings grouped by key themes starting with the major challenges and opportunities uncovered within the DH data reuse infrastructure: training and education needs, shifting roles in the digital research infrastructure, improvements to data discoverability, curation networks to boost expertise and cross collaboration in data stewardship and preservation, and data sharing pathways to data reuse. I conclude with a discussion on culture change.

1. Education (5.1.1)
2. Shifting Roles in the Research Infrastructure (5.1.2)
3. Data Stewardship, Curation and Long-term Preservation (5.1.3)
4. Improvements to Data Discoverability (5.1.4)

5. Data Sharing Pathways to Data Reuse (5.1.5)

6. Culture Change (5.1.6)

Two of the biggest findings in this project were in RDM and data enrichment. First, individual scholars need better, earlier education of the importance of RDM training at early stages of academic careers in order to enable open science and drive home the importance of the role of data sharing as an enabler of open science. Second, digital humanists make undervalued contributions to data ecosystem by enriching datasets after reuse.

There are also some noteworthy trends in participant responses that call for collaboration. I frame these as strengths in the form of recommendations, in particularly in the formation of networks of data stewardship, and call for a bottom-up, socio-technical approach to the research and academic infrastructure. I make note of one caveat: it is difficult to make international recommendations without knowing the cultural nuances and needs of communities and regions. Therefore, I make them from my vantage point in Canada.

5.1.1. EDUCATION

There are currently gaps in RDM training and data literacy, especially within HSS (Garstki, 2022). But as my interviews with scholars showed, digital humanists excel at learning new skills on their own. The majority of participants in this study had training needs, saying that they upskill by necessity according to the type and format of data they wish to work with. This represents both a challenge and an opportunity for DH

reusers. On one hand, training may benefit the scholar (professional skills for the job market, learning for personal interest). Conversely, training also represents spending precious resources – time, money, and support.

Although research institutions and funding partners "pursue somewhat different goals" in mandating open data, "pedagogy knits them together" (Garwood & Poole, 2019, p. 567). In other words, education, pedagogy, training can help link the tasks of scholars to the digital research infrastructure and eventually, the goals of open science. However, participants in this study claimed they either were not at all interested in training in research best practices or were interested but felt they had no time for training. This needs to be considered when designing learning opportunities for researchers in DH. If university research leadership offices need to look at the big picture of digital research infrastructure, and act in response to the scholars' needs (e.g., support for more training) perhaps more time could be freed up for pedagogical pursuits. Participants noted that it was important to them that RDM-related data literacy skills begin at the undergraduate and early graduate level. There may be limited opportunities for students to learn about RDM, depending on their university or geographical context. New scholars may not know fundamental RDM practices or even where to learn. It is important to treat these scholars as experts at learning new skills on their own and meet them where they are. Libraries and repositories should practice regular outreach to reach as many communities of scholars as possible – this is how RDM practices become embedded and normalized within academic disciplines.

More than one participant said they could teach an RDM course, and those who did not, said they would have no time for RDM training. When it came to RDM, one participant said it was a “very important topic” that was not being taught enough to PhD students. Another said DH scholars should be encouraged to “think of themselves as” data stewards. The data reuse infrastructure for DH and other humanities scholars is marred by socio-cultural, technological, political, organisational, economic, and legal barriers stand in the way of opening up science. Clearly linking open science to RDM would help researchers see the bigger picture and situate themselves in the lifecycle. At the University of British Columbia, an undergraduate Open Science instructional program has been piloted, and another one in the U.S. for data stewardship (see p. 116).

When designing RDM training and data literacy interventions for scholars, research supporters should understand and work in harmony with digital humanists’ culture of self-learning. Learning from within one’s own discipline via peer-to-peer education can be an effective approach given the unique data use and reuse experiences scholars share. This type of pedagogical approach normalizes data sharing and data reuse practices, as Yoon and Kim remark “education by members of one's own research community” enforce “a subjective community norm of data reuse” (2017, p. 231).

Data literacy for scholars should begin early in their academic career, at the undergraduate and postgraduate level. Data literacy should begin early in scholars’ careers.

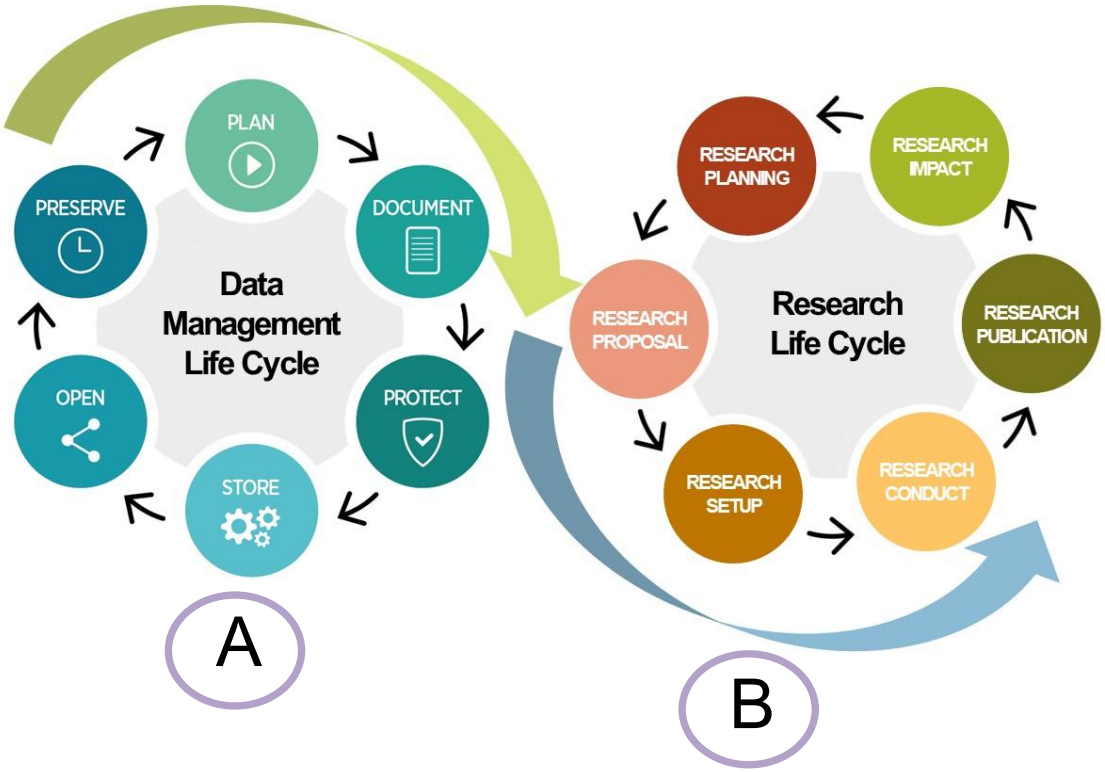
Recommendation 1: Address data literacy at several levels: peer-to-peer, self-paced/autonomous, and at the undergraduate and graduate level

It is also important to demonstrate to new scholars how they fit into the RDM ecosystem, show them how to position themselves within it, and situate this knowledge within practices of open science. Cox and Tam (2018) suggest researchers do not necessarily see themselves in the research data lifecycle, and participants in this study certainly showed this to be true. In Figure 6, it is clear that project phases of the RDM lifecycle (A), reveal separate but complimentary project phases of academic research (B).

Recommendation 2: Help Researchers See Themselves as Part of the Research Data Lifecycle.

Figure 6

Research Life Cycle and Data Management Life Cycle



Note. By Tanja Lindholm, Mikko Ojanen & Liisa Siipilehto/University of Helsinki. The research life cycle and data management life cycle shown as separate but overlapping processes. From "What is research data management (RDM)?" [Image], *THINK OPEN*, September 3, 2020, https://blogs.helsinki.fi/thinkopen/know-your-data-rdm-series-1/

5.1.2. Shifting Roles in the Research Infrastructure

Institutional and public policy with the humanities in mind is needed to ensure the long-term handling of research data over time. The slow change towards a culture of

open science in academia is almost certainly the biggest high-level, structural barrier to data reuse in DH. As this study demonstrated, there are numerous improvements that can be made at the policy, individual, and institutional level to better enable open science.

As research policy makers and funders continue to mandate data deposit, they will have to look more closely at the sustainability of this ask in regard to compliance. compliance monitoring to a data deposit requirement to funding has not yet been studied, but we can perhaps see the effects of compliance monitoring through another open science mechanism – open access publishing (OA). Larivière and Sugimoto (2018) found that compliance increased when funding policies required scholars to publish research conducted with public funds. Clear policies and monitoring led to authors being more likely to publish their articles under an open access agreement with the publisher. If policy makers and funders wish to see greater data reuse and data sharing, they will need to develop clear data deposit policies and monitoring if they wish to see the uptake seen with OA in journal article publishing.

Another open science strategy might be for librarians, data stewards and curators to act as research infrastructure mediators to facilitate a bottom-up approach. A bottom-up approach, socio-technical approach ensures that research infrastructure scale-up is not only be responsive to researcher needs, but is supported and mediated by librarians, data stewards and curators. Advocating for researcher needs from the bottom (researchers) to the top (funders) could enable buy-in of training, better compliance to open science policies, and may enhance collaboration among scholars

across disciplines. A researcher-centric, bottom-up approach is discussed in the recommendation under section 5.1.3. (Data Stewardship, Curation and Long-term Data Preservation).

Scholars appear to be, by and large, mostly left to figure out a place to manage their data. Yet Borgman argues that the ability to release, share, and reuse data (and their representations) “depends upon the availability of appropriate knowledge infrastructures to do so” (2015, p. 206). Borgman’s conception of knowledge infrastructure are relevant in this study because they suggest they are *indivisible* from scholars’ data reuse practices. Thinking about data infrastructure and research data services can help us see how they impact the experience of researchers. Research organizations are socio-technical – enmeshed in policies, people and the technical. As such, they can only be understood and improved upon by thinking of those three pillars as interdependent parts of a system: “The socio-technical approach considers the individual and institutional or communal creators, stakeholders, audiences and their interaction with the technologies of archiving, digitisation, storage, access, display, navigation etc. in a complex web” (Bow, 2019, p. 108).

An analysis of the DH scholars that served as interview participants in this study show the socio-technical aspects of this knowledge infrastructure – data reusers are part of a holistic system and their experience should be reflected in how data infrastructure systems are built, designed, improved upon, and maintained. There are still unanswered policy questions and infrastructure challenges regarding research data, reuse and open science that need to be better understood in order to integrate a culture

of data reuse (and RDM) into DH (Pawlicka-Deger, 2021). Scholars in engineering have suggested that adopting a socio-technical approach may lead to systems that are more acceptable to end users and deliver better value, and more sensitisation and awareness of the pathways to holistic and sustainable change (Baxter & Sommerville, 2011).

This research project has shown that DH scholars are looking for data to reuse for research purposes, and note that there is more and more data collected, stored, and reused in an academic context. These scholars want to focus on their research goals, not managing the data that could potentially underpin their new research findings. One participant said that RDM “shouldn’t be an afterthought.” While another participant said his work responsibilities included being a data manager, many others thought RDM was not their job. One participant was open to the idea of being her own data steward for the data she worked with, but in practice, she felt this was not yet a reality. Thus, more support is required to effectively manage reused data stemming from academic research. Research has already pointed to the value of and need for specialised support professionals in RDM – such as information managers, data librarians, data curators and stewards – to keep up with the pace of ever-evolving technologies and shifting research data policy landscapes (Berman, 2017; Federer, 2018; Ohaji, Chawner & Yoong, 2019; Poole & Garwood, 2018; Tenopir, et al., 2014). Interested parties working in research peripheries in academia could also be involved to help with RDM training and awareness building of open science – for example, data champions, DH scholars, graduate students.

Libraries and librarians can do a lot to increase awareness in RDM and in the services they offer in RDM support. Institutional leaders should place more value on training in this domain, coupled with increased funding and opportunities for librarians with archival, collection management, preservation, domain, and scholarly communications expertise. Employing more librarians with data specialization and increasing data support training for librarians would go far in improving the data reuse experience of DH scholars overall. Likewise, increased support should be coupled with increases in funding for the people aspect of the infrastructure that support the academic research infrastructure.

Recommendation 3: Professionalize and Promote RDM support roles including in Data Governance, Curation and Stewardship

5.1.3. Data Stewardship, Curation and Long-term Data Preservation

Curation networks are another potential pathway to connecting scholars to RDM and the research data lifecycle. The digital research infrastructure is composed of many actors – libraries and their institutions, private and government funders, and research support organizations. It is the responsibility of all actors in the research infrastructure to effectively manage data, from all stages of research projects to all stages of the data management lifecycle. As mentioned earlier, there is room and immense value in advocating for continued professionalization of RDM roles, such as data stewards or data librarians (Jetten et al., 2021). Curation of data, or data stewardship in Europe, is

one mechanism by which repositories and other infrastructure can take up the researcher-centric, bottom-up approach and ensure their success and buy-in.

Research support organizations and academic libraries provide the services needed to mediate technical infrastructures, such as with data repositories. Repositories have grown rapidly in recent years, to keep pace with the growth and use of data needed to conduct scientific research. Because of the volume of data, as well as the software and hardware needed to power processing of these data, data repositories are supported by the larger technical infrastructure of high-performance computing (HPC). As more and more heritage and cultural institutions digitize tangible and physical material, these data are being deposited into publicly searchable data repositories (i.e., Kaggle, government sites, web archives, museums, libraries). And with DH, it is possible that scholars would want to access cultural and heritage data, which may come from either institutional data repositories or publicly searchable repositories. These digital repositories spaces are in constant flux, and as storage demands increase, more and more energy go towards preserving data. These spaces include data warehouses and repositories, data lakes, commercial and public clouds, and super computers like McGill University's Béluga, or the University of Toronto's Niagara. Data lakes are essentially a way to park raw data – structured and unstructured, organized, and curated (Huchard et al., 2020). These infrastructures sometimes operate in silos and have been normalized within certain disciplines, especially in ecology and ocean science. For example, the Canadian Integrated Ocean Observing System (CIOOS) provides a research data catalogue, deposit and curation system that has been quite

successful for its users. This network was created because of users (researchers) that direly needed a way to catalogue, curate and store their data in order to maximize their time and efficiency. CIOOS is a good example of what is lacking in other disciplines, and equitable, organized, shared platform that responds to user needs and provides access to data infrastructures. Internationally, there has been an emergence of curation networks, or the embedded data stewardship model[5]. For ease of understanding, data curation refers to external data professionals working with researchers and their data, and data stewardship will refer to an embedded subject matter expert that works with researchers and their data.

Curation networks “collaboratively address shared and difficult challenges in data curation, to promote ethical data sharing, and to foster open science” (Carlson et al., 2023, p. 12). There is evidence that shared data stewardship models or data curation networks are a way to streamline data sharing and data infrastructures, while maximizing resources and collaboration. Operating since 2016 with a grant from Alfred P. Sloan Foundation, the U.S.-based Data Curation Network (DCN) is the first data curation network to exist. Similar international efforts have since followed suit. DCN is made up of institutional and non-profit data repositories who work to advance open research by making data more ethical, reusable, and understandable (Carlson et al., 2023). Other efforts are under way in Australia (for example the Australian Research

[5] Data stewardship is more often seen in European research contexts, and data curation in North America.

Data Commons) and networks are being conceived of in European countries (Ireland, Sweden, Netherlands, Germany, Norway, Switzerland, Italy). In the U.S., the Council on Library and Information Resources recently started offering data and software curation post doctorate fellowships. An American university in Virginia piloted a student "data champions" network aimed at enhancing research data services, augmenting openness in research outputs, and extending the reach of the work beyond the researcher's community (Briganti et al., 2020). There is also notable advocacy work happening with non-profits working on and in the research infrastructure. U.S.-based Ithaka S-R publishes influential position papers to leverage and build data infrastructures that advance open science (see for example, the 2023 Ruediger & MacDougall article: Are the Humanities Ready for Data Sharing?). Research Data Alliance, an international open data advocacy organization, brings data stewards and disciplinary experts together to do similar advocacy and community building. At DCN created a "radical interdependence" model as a framework for how curators in a distributed institutions work together yet independently to work through data curation challenges while sharing workflows, techniques and administrative or funding models (Carlson et al., 2023).

The consequence of not thinking through how data will be preserved for the long term properly may mean that datasets are lost or unfindable and leads to research waste. Repositories were an important theme that emerged during interviews with scholars. While many participants said they used GitHub, their data was spread out onto many repositories – private, commercial, and institutional. One scholar worked for an employer who provided a website to store data, which is risky for the scholar

because of possible loss of data, issues with IP, and missed opportunities for employment or collaboration. Improper storing of research data may not be findable – repositories usually have the ability to mint DOIs in order to persistently link digital objects to publications, affiliations, and scholar. Linking data to academic articles requires forethought. Most of the scholars who worked at institutions (7 out of 12) said they did not want to commit to storing their data in a single location. Duplication of data is harmful to the integrity and authority of scientific research. If duplicates exist, how will anyone know which version is the version of record – the authoritative and complete version?

More robust preservation and curation policies at institutions, research support organizations, and funders would help ease the environmental and monetary burden of keeping all research data in storage. Data should instead be selectively appraised, and those data chosen for long term preservation should be accompanied with guidance for reproduction of large datasets can be rerun using carefully detailed instructions in data documentation. Not all data need to be preserved for the long term, it depends on factors such as the size, ease of reproducibility, fragility of formats, and long-term value of datasets. Dedicated curation staff and preservation appraisal models can help mitigate researchers' estimated versus actual storage needs. The lifespan of research data may end when they are no longer used, become irrelevant, or when new, better data supersede them (Stigler. (2019). There are monetary and environmental costs to simply keeping all research data as well, therefore research data should be appraised and carefully selected for long-term preservation based on a number of factors. Thus,

the curation process should also integrate a preservation analysis that looks at the quality of the dataset, its format for accessing it in the future, preservation objective and whether it has a designated user community that will understand it and gain knowledge from it (Conway et al., 2009). When data preservationists decide not to preserve data for the long term, modelling outputs, metadata, and instruction for how to reproduce the dataset can be included in data documentation. This is in line with the archival institutional policies of many heritage and cultural collection stewards across the world (Duranti, 2005).

> Data curation networks increase efficiencies and collaboration in data sharing and data reuse. Ensure they are informed by, modelled after, and grounded in international principles and best practices. (For example, OCAP – Ownership, Control, Access and Possession Principles, the FAIR Principles, TRUST Principles for digital repositories, and the CARE[6] Principles for Indigenous Data Governance).
>
> **Recommendation 4: Support and Invest in the Creation of Curation Networks to Boost Expertise and Cross Collaboration in Data Stewardship and Preservation.**

5.1.4. IMPROVEMENTS TO DATA DISCOVERABILITY

[6] An international model of indigenous data governance, which stands for **C**ollective Benefit, **A**uthority to Control, **R**esponsibility, **E**thics.

Participants in this study spoke of a wide range of varying and differing ways they discovered data, and how they practiced RDM by organizing, storing, and sharing data on private or public websites and repositories.

This study noted a unique feature of data reuse among DH scholars. Participants mentioned that they draw research project data and ideas from multiple sources – not just conventional academic research data sources. Some of the sources for data, therefore, comes not only from institutions such as the British Library, the Rijksmuseum, Kaggle, Government data portals or Zenodo, Kaggle and Google Datasets. One participant once pulled birth stories from Reddit.com; another pulled together scattered parts of a Mexican author's manuscripts from various web archives. These paths to sources of data show that DH truly straddles the boundary between research data proper and the murkier less organized world of web archives. When it came to how they practiced RDM, how they organize, store and share data is interesting because it points to how little consistency there are in the ways in which RDM is practiced.

How data is discovered was a unique finding of this study as well and can be intrinsically Carol Kuhlthau's work in information seeking behaviours. In a large-scale meta-analysis of the data reuse process literature, secondary data reuse seeking behaviours were identified, building on Kuhlthau's 1991 information search process theory (Kuhlthau,1991; Wang et al, 2021). The process of reusing data, from the user's perspectives, is four-fold: 1 – make a decision, 2 – discover and acquire data, 3 – understand and choose data, and 4 – process and reuse data (Wang et al, 2021). In this thesis project, participants easily made decision to reuse, but came across barriers

at the second stage – to discover and acquire. Data acquisition was a consistent and recurrent theme in participants' response. This is an important finding of this project – data reuse is dependent on not only data sharing practices of scholars, but on the capacity for data reusers to discover the data. With the data reuse era already upon us and given that data reusers in DH appear to build their work on other researchers' findings, bettering the discovery experience of these scholars is also key to improving their experience.

Repositories play an increasingly crucial role in the digital research infrastructure and open science environment (Barsky, E. et al., 2018), and discovery repositories that use federated searching have been shown to improve discoverability of digital objects (Turp, et al., 2020; Devarakonda et al., 2011; Wang & Mi, 2012). Federated repositories are a collection of repositories, which harvests and indexes metadata records to make research data available from a single web portal. In Canada, the federated discovery service Lunaris indexes repositories from federal, provincial and city government portals like the Government of Nova Scotia Open Data Portal, the Canadian Space Agency Open Data, STEM labs such as the Global Water Futures project, as well as Borealis, the data deposit service installed at a majority of Canadian universities. Aside from Dataverse collections from Canadian universities, there are low amounts of humanities data available from this federated repository. This is a missed opportunity to enable discovery of humanities data for scholars globally. A federated repository "serves a diversity of needs, drives traffic to host repositories, and helps break down data and disciplinary silos" (Ontario Council of University Libraries Scholars Portal et al., 2019, p.

4). Federated repositories are not unlike a repository *directory*, such as the re3data initiative. Re3data provides a listing of repositories sortable by discipline, however, the discovery and searching of each repository individually is still incumbent on the researcher. Individually searching in a consistent manner in each repository – like in a search strategy for a scoping review, for example – may be a difficult and long process for scholars. Conversely, this again shows the value of the data librarian or data research support professional in aiding in this phase of research.

Indexing singular cultural and heritage collections repositories into federated repositories would improve the discovery of these cultural and heritage data collections for researchers. There is evidence that generalist repositories break down data and disciplinary silos (Ontario Council of University Libraries Scholars Portal et al., 2019). And while domain-specific or disciplinary repositories have their place, indexing and linking multiple data archiving systems together in a generalist repository may allow for more social scholarship, greater findability, and interdisciplinary crossover. Setting up a federated search of repositories, while possible, is technically difficult. Institutions should take on this infrastructure challenge to improve data discovery.

Federated repositories can break down data and disciplinary silos.

Recommendation 5: Ease Difficulties in Data Discovery by Linking Discovery Mechanisms Such as a Federated Data Discovery Service

5.1.5. Data Sharing Pathways to Data Reuse

As mentioned in 4.4.5. (Data Reuse Leads to More Reuse and Richer Data), one of the enablers of data reuse is the possibility for scholars to enrich datasets, patching together datasets to make a larger dataset. Ultimately, data authorship can lead to a sense of pride and ownership over data created, and if the enriched dataset is shared and published, the scholar may be incentivized to continue to contribute to the knowledge infrastructure in this way. In the academic climate of “publish or perish”, data authorship can be another path to obtaining credit and receiving personal benefits (such as career incentives in academia). The alternative path to scholarly credit would relieve the pressure of publishing for the scholar, rather than counting solely on contributions of articles in journal publications (Alperin et al., 2020; Gregory et al., 2023).

Every participant in the study said that they believed in the benefits of data sharing, yet their practices did not all match this belief. Part of the problem is in defining at which point in the research of the scholar should RDM be engaged. For example, when asked about RDM, Dr. Daniel stated that he thinks of RDM as being someone else’s job. Conversely, when asked about data sharing, he stated that he supports open science and believes in data sharing. But a belief does not always translate into an action. Ultimately, data authorship could help enable increased citations, IP, and credit for contributions to the knowledge infrastructure.

If the pathways to data sharing rewards for researchers was easier, could rates of data increase across a diversity of disciplines, even in the humanities? Is it possible that academia has a data sharing problem, rather than a data reuse problem? While researchers are willing to share their data, few actually release their data (Tenopir et al.

2011; Wallis, Rolando and Borgman, 2013). In looking at data reuse motivations, the DH scholars in this study found it difficult to find data to reuse in their research. More data sharing by data creators would have augmented the relevance of discovery results for these DH researchers. There are benefits, challenges, personal motivations, as well as policy and technical reasons why sharing data is not more widespread. Since sharing research data is associated with increased citation rates (Piwowar, Day & Fridsma, 2007; Tenopir et al., 2011), it is worth continuing to probe what other barriers there are for not sharing data.

How data sharing is perceived needs improvement. Researchers in the U.S. propose stronger language around data dissemination that help signal an incentive. Atici et al. suggest using language like “data publication” to improve reuse of research data (2013). “Publication” carries more weight and is more reflective of the effort required to share data. It also unambiguously signals that data publication can and should be a recognized as academic labour. Terms like data publication are more accurately descriptive than terms like data sharing, data deposit, data release, or data archiving. Publishing data has more cultural and scholarly cachet, and may drive the push towards rewarding data creation, publishing, and citation. Devriendt et al., 2021 suggest that there are equally not enough incentives to incite scholars to publish data in the biomedical sciences. To incentivize data sharing, the humanities need more education and awareness about alternative recognition mechanisms, data publication and attribution for derived and sourced datasets. Lowenberg, et al.’s Code of Practice for Research Data Usage Metrics is a good example of a tool that could help data

repositories evaluate how they could potentially produce and integrate a consistent, comparable, and credible metrics system for published research data (2018).

Recommendation 6: Both researchers and repository service providers need more awareness of alternative recognition mechanisms such as data publication, attribution for derived and sourced datasets to incentivize data sharing.

5.1.6. CULTURE CHANGE

Digital humanists can act as data champions, to not only teach and demonstrate data literary to researchers on campus and in other seemingly disparate communities of data users (librarians, commercial and public data infrastructure providers, the public). As a field, DH has often looked back while looking forward – using new tools to probe and analyse history with new eyes. And as the "data as collections" model shows, DH can take the position of a "more data-focused look" as a case study discipline. DH is positioned to lead culture change in research data management, open science pathways in academia, and data reuse.

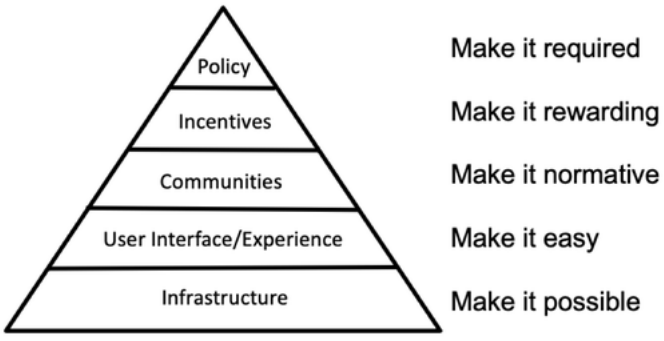
Culture change will take purposeful collaboration. The recently published *Data Primer: Making Digital Humanities Research Data Public* book in Canada offers support for DH scholars, and humanities scholars in general. Librarians and DH researchers worked together to make recommendations for scholars at all stages of research projects, including data flow models for handling issues like consent, data collection and processing, IP, analysis, sharing and publishing (Tayler et al., 2022, "How Do You Work

Through the Data Curation Management Process?"). More of these types of librarian-researcher collaborations are needed across fields.

The best chance at culture change will likely resemble Nosek's model of culture change diagram (Figure 7). It shows that, for change to become normative, it will need to shift towards a model of where there is a sense of both collective responsibility (the infrastructure) and individual responsibility (e.g., disciplinary communities). Researcher-centric decisions in the infrastructure, easy to use user interfaces, career rewards and incentives and policies are also of equal importance. Bearing these elements in mind – while taking a good look at the data resulting from interviews with scholars in this very study – a culture shift can only be attained if DH scholars are engaged in data strategies, and their experience is considered and understood.

Figure 7

Strategy for Culture Change



Note. By Brian Nosek. A Culture Change pyramid with different levels of action is shown. "Strategy for Culture Change," [Image], June 11, 2019, *Center for Open Science* blog, https://www.cos.io/blog/strategy-for-culture-change

Nonetheless, policy should not be the only impetus to practicing open science with data, as noted by Pasquetto et al. (2017). What is more, it is important to remember that there are significant technical, investment and data access barriers – "the promised benefits of open access to research data lie in the ability to reuse data" (2017, p. 7). In the U.S., all federally funded researchers will need to share underlying data at time of publication starting in 2026 (OSTP, 2023). The National Library of Medicine at the NIH recently positioned the importance of advancing cultural and technical aspects of the data economy in research, as they begin to implement the recommendations: "We cannot survive the data revolution without significant federal investment. We need to build the intellectual infrastructure for discovery.... We must create technical and cultural changes to make this happen" (Stall & Martone, 2020, para. 4). This statement from the NIH links the technical and cultural advancement of RDM as the key to establishing a culture of open science, but that can easily lose support without proper investment in the infrastructure needed to support these changes (Stall, Shelley et al., 2020; Stall & Martone, 2020).

This study took a necessary close look at the socio-technical underpinnings of the research infrastructure, attempting to probe the social and the technical not just in isolation, but to "examine the interactions that occur where the two intersect" (Gregory et al., 2020b, p. 472). Collaborative change is central to improving the experience of

data reuse among DH scholars in the research infrastructure. We know that collaboration is not only a hallmark of DH (Nowviskie, 2014), it is “a community of practice” and “the sine qua non of DH education” (Garwood & Poole, 2019, p. 551). Its interdisciplinarity with other fields and its distinctive approach to data methodology show that collaboration is both a part of and a condition of DH (Burdick et al., 2016). DH may be uniquely positioned to be an early indicator of open data successes, failures, and pressure points. In her 2014 talk “Digital Humanities in the Anthropocene”, Nowviskie proposes minimal computing and asks DH to think about the intersection of their data use with “the ephemeral, preservation, archiving” of technical resources (2014, p. i4). This research showed that DH scholars use library collections as data. DH – as a large, distributed disciplinary community – can act as a use case scenario for libraries to change the current paradigm – from a “push” to a “push pull” paradigm (Padilla, 2016). Although a push paradigm appears to be the current de facto approach that governs data access strategy in many academic libraries, Padilla argues that institutional libraries will need to “consider how to incentivize pulling humanities data back into the collection once it has been used for a given application” (2016, para. 28). A push-pull approach to data collection development supports reproducibility (Faniel & Zimmerman, 2011), technical infrastructure and “data preparation practices with a growing emphasis on the value of data reuse” (2016, para. 28). The approach proposed by Padilla would ensure proper data governance that include provenance tracking and proper attribution based on a dataset that has been derived or sourced from one or more datasets, enriched, and redeposited into the collection as a new data output. There are multiple

pedagogical and research benefits from communities (libraries data creators and data users, and peer reviewers) in this approach (Padilla, 2016). This push pull paradigm specifically concerns improving access to data from the source and charting its path all the way to its first, secondary and subsequent uses. Look at the lifecycle of individual dataset helps the data holder understand what elements of that dataset makes it popular or “in demand”; or can then either improve its metadata and documentation, or move the data into cold storage if it is underused. This proposed access paradigm is in alignment to the findings of this study, which propose a bottom-up, researcher centric approach to culture change in socio-technical infrastructure research organizations.

5.2. Further Research

The data reuse landscape is still new, as well as incredibly dynamic and evolving. Some of DH’s future research areas might look at affording equity to DH scholars working with data, as well as the environmental risk and harm in storage of data. Nowviskie asks the DH community to consider the “environmental and human costs of DH. From supporting the Big Data Industry, to the carbon footprint and price tag of big conferences, digital humanists need to ask what they could change right now, ‘or grow to be’” (2014, p. i13). There are unequal parities to be considered in the field, as “the so-called global in humanities technology is not equally distributed” (p. i12). Data sharing, archiving, discovery, and reuse may not be accessible to all scholars and digital humanists, and greater attention to matters of sustainability, accessibility and minimal computing need to be paid in order for all scholars to be able to participate with

equity in the research data economy that lives within digital data archives, essential pieces in the current academic knowledge infrastructure (Borgman, Scharnhorst & Golshan, 2019).

As mentioned earlier, research policy makers and funders continue to mandate data deposit yet have not assessed the sustainability of this requirement from a compliance perspective. There is another aspect of the sustainability of this request – its ecological impact. There are “frictions” between data and energy infrastructures to consider in the future of research data sharing and reuse (Velkova, 2019). However, the future of storing research data is uncertain. Increases in data storage must keep pace with environmental capacity and consider its footprint. Data storage is incredibly burdensome to power grids, and there are high costs to the use and manufacturing of cables, carbon, heat, hard disks, and data tower cooling systems (Monserrate, 2022; Zakarya, 2018). More research is needed to improve research data infrastructure and its “very material consequences” (Hogan, 2015, p. 7). Improving data repository storage systems would have major repercussions to climate change, so while improvements to the data reuse economy might benefit the knowledge infrastructure and the ease of its reuse by scholars from all disciplines, unchecked data storage growth is unsustainable. Some of the solution to this storage and environmental concern that have been proposed are data modelling, and robust preservation and curation policies at data-hosting institutions. Depending on the dataset, modelling outputs tend to be better for long term storage for many reasons such as environmental harm (maintenance and cooling of computer towers for example).

Interdisciplinary collaborations drive data sharing. One participant mentioned being part of the GLAM Workbench community for learning to use and wrangle GLAM data. In biology, a platform called *napari* is a community-driven and patient centred research tool used for browsing, annotating, and analyzing large multi-dimensional images. Tools like *napari* and the GLAM Workbench are examples of collaborative science. What would DH and other fields in HSS look like if the technology could catch up with the needs of researchers? Technological capabilities enable collaboration, why not build large-scale collaborative tools to annotate cultural heritage digital and representational artifacts? As more physical artefacts become digitized, the reality is at our doorstep.

And finally, the model offered by the First Nations Principles of OCAP (Ownership, Control, Access and Possession) is such an interesting model to use a framework for data governance: “OCAP asserts that First Nations alone have control over data collection processes in their communities, and that they own and control how this information can be stored, interpreted, used, or shared” (First Nations Information Governance Centre, n.d., para. 3). Carroll et al. posits that a good data governance strategy should always include defined cultural metadata, and provenance should be recorded in order to achieve unity among principles of data governance such as FAIR and CARE (2021). Sengupta suggests that data governance could be a responsibility headed by data curators or librarians to help navigate governance and ownership issues due to their expertise in the management of data, while subject specialists would signal when they are blind spots thanks to their data governance expertise (2022).

CHAPTER 6: CONCLUSION

The thesis set out to find out how digital humanists use, find and reuse data and what barriers and enablers to data reuse exist among digital humanists. How do digital humanists use, find and reuse data (RQ1)? It turns out that the actual locations where DH scholars look for data demonstrate that they are not only looking for “research data proper” in conventional places (like institutional repositories) but they are looking at GLAM, commercial and web archives to source data for academic reuse. DH scholars find data to reuse in a plethora of places such as institutional repositories and academic library databases, and GLAM archives (physical or digital) such as the Library of Congress, the Rijksmuseum, and the British Library. Scholars use generalist repositories like Dryad, Dataverse, Figshare, Zenodo, as well as discipline-specific repositories like the Commons Open Repository Exchange. They also use commercial sources of data archives, including commercial data hosting services such as GitHub, Kaggle or Google. DH scholars also source from web archives such as the Internet Archive and the WayBack Machine, Wikipedia, Wikidata, public discussion forums, Reddit.com, GoodReads.com and government data portals. Many of the scholars also use APIs to mine and harvest datasets to extract data to use in a research project. Most interestingly, digital humanists not only combine and enrich existing datasets – they will also pivot or discard research questions if they cannot find or access the right dataset for their purposes.

In RQ2, the study explored the barriers and enablers to data reuse among digital humanists. It found that there are almost as many barriers (9) than there are enablers (8) in data reuse for this population. However, the barriers may be harder to address because there are more structural barriers, and these are the types of barriers that are the hardest to unravel. Non-structural barriers to scholars reusing research data included technical training gaps, labour intensity and lack of time and resources. This research also found that, among the enabling factors to reusing data, digital humanists are self-taught, autonomous workers that can also be collaborative in their approach to using and searching for data. Further, the ways in which they discover, reuse, and use data are complicated by deeply individual searching preferences and technical heuristics. Although their use, reuse and discovery heuristics are often a solitary act as they perform as researchers, searching for data can often be an opportunity for collaboration with data producers, data curators, and librarians who manage or hold this data.

The remaining barriers to scholars reuse of data in the research infrastructure can be summed up as a culture clash between scholars' needs, and the requirements and limitations of funders and institutional leadership, resulting in an academic culture that is not optimized to meet the needs of DH researchers. Data reuse practices are deeply embedded in research culture, and therefore future priorities cannot be divorced from their cultural makeup. A bottom-up approach would be helpful in addressing challenges with policies that prioritizes the researchers' needs and be more conducive to open science.

One of the barriers to data reuse highlighted in the findings of this study show that inconsistent data practices and low rates of data sharing among data producing researchers could be addressed by better RDM training at the undergraduate and graduate level in universities, including continued data literacy throughout the career of the scholar. Better recognition of data publishing (collecting and then sharing datasets), better promotion and rewarding of alternative recognition mechanisms, as well as robust policies in data citation could help raise the profile of data producing researchers who publish datasets and thus mitigate low rates of sharing across all disciplines. The research culture could further be enhanced at the infrastructure level with improvements to data discovery systems: support of and investment in research data curation networks that would see data stewards govern data through its lifecycle as data experts, alongside subject specialists, and data owners. Data governance models such as CARE, OCAP and TRUST principles should be explored to see how their data governance frameworks could be incorporated into socio-technical organizations such as libraries, museums, and repositories. Difficulties in discovering and using data fit for reuse could be lessened by linking cultural, heritage and web archive data discovery services through a repository indexing mechanism such as federated data discovery services.

Participants revealed a number of enabling factors in data reuse, and chief among them were a collaborative openness to contacting scholars to obtain data they wish to reuse, and a spirit and enthusiasm to learning new technical skills. Data reusers and data producers also have to understand their roles and responsibilities as

researchers in a complex new data age – and this may mean they embed research workflows into the research data lifecycle workflows for the sake of supporting a more open way of doing science. To get there, there are socio-technical pathways to improving how DH scholars reuse data that would best be addressed with a bottom-up and researcher-centric approach, as well as improvements to the technical and cultural research infrastructure, data literacy and policy. Technical upskilling of librarians and professionalization of data services experts is tantamount to establishing sustainable, regionally specific, networked data stewardship and governance.

In order to normalize data reuse and sharing practices of scholars, more qualitative and disciplinary-specific research is needed to better understand the attitudes, practices, and motivations for sharing and reusing data. The qualitative study scope of digital humanists in this study ultimately provided a case study for inquiry into the practices, challenges, and opportunity in research data reuse. Alongside a better understanding of their practices, scholars need more education about the layers involved in data reuse practices as the knowledge infrastructure begins to edge closer to open science principles. As more physical artefacts and their digital representations become digitized, and as funders begin to require data to be deposited for public accessibility, an open science approach to data in academia is closely approaching. Yet this reality will remain merely a vision of the future without sustainable structural improvements to the knowledge infrastructure that starts from the bottom-up and is informed and governed by the researchers and librarians that use and manage the system the most. This community of users should form data curation networks to bring

their subject expertise and work in tandem with institutions, journals, policymakers, and funders if they want to see changes in the culture, including any significant rises in data sharing, or enhanced data practices in DH and the humanities. However, those with influence at the top of the infrastructure must commit to working with the community to see improvements to the research ecosystem, and this commitment must be coupled with shared understanding, the recognition of collaborative data curation networks as well as investments in time and money.

Infrastructural change that is sustainable to the users that make up the research data ecosystem likely needs to come from the bottom-up in order to be meaningful. Pawlicka-Deger has stated that data reuse in DH needs to be situated within a research culture shift: “it is essential to take the bottom-up approach and present open research data as a part of a new research culture, rather than something that is only necessary in order to fulfil institutional and funding requirements” (2021, p. 55). The three-legged stool metaphor (see section 2.1) highlights the equal importance of technology, organization, and resources. This infrastructure is made up of three pillars that “hold” up the system (in this case the data reuse or RDM research culture) – technology (repositories, data lakes, discovery layers), organization (institutions) and resources (people). Each is essential in holding up the research culture, and when one fails – the entire structure can crumble. Metaphors such as the data lifecycle, the three-legged stool, or the research data ecosystem, are essentially a visual depiction of workflows with a suggested order of operations. The socio-technical approach is about showing

scholars and infrastructure leaders how these pieces flow together – and that is an essential key to culture change.

Collaboration with technical support and research support staff is also crucial to culture change. A solid socio-technical approach to work culture organization in academic libraries and research centres improves the institution's research reputation, (along with data management and information management implementation) because research support staff are actors in the ecosystem that are able to gain insight into storage capacity needs versus reality. These backstage actors in the ecosystem will need to be engaged with and consulted in order to inform discussions with funders, journals and policymakers requiring data deposit. Policymakers, journals, and funders need to hear and listen to research data support staff; just as support staff need to listen to the needs of the researchers.

Researchers creating datasets also need to understand their role and responsibility in complying with data deposit requirements and best practices as best they can. If research data concerns sensitive, proprietary, or human data – as humanities data often does – researchers should find ways to compromise. For example, instead of simply *not* depositing or sharing restricted or closed data, researchers could create and deposit metadata records into an online catalogue containing rich descriptors of the research. Depositing documentation and metadata files, alongside contact information, would enable researchers looking to reuse their data to enter into communication together and potentially make a case for the reuse of the restricted data (with justification). Data-creating researchers then have control over

the use of their data, they know about the reuse, and will have not only complied with a data deposit best practice or requirement, but they will have made an individual contribution to enabling open science. Early and fulsome RDM training and data management planning enters into this workflow as well – as researchers embark on new research projects with humans, they must make their intention to share data clear to REB offices and participants before data collection begins.

This thesis has tried to show that within these needed culture shifts, it is also just as important to raise, augment and place value on the voices of DH scholars' and their experience in data reuse. The findings in this research show the importance of research support staff, data stewards, curators, and librarians as essential in strategic conversations about digital collections – a bottom-up approach to uniting disparate research cultures.

Finally, it is striking to see that digital humanists in this study improve the data infrastructure in each of their local spheres, simply by augmenting, combining and enriching datasets they have reused. However, it is also concerning that the availability of data threatens and limits science when one considers that scholars adjust research questions, or discard entire research projects, according to the data that is accessible or available to them.

References

Abella, A., Ortiz-de-Urbina-Criado, M., & De-Pablos-Heredero, C. (2019). The process of open data publication and reuse: The Process of Open Data Publication and Reuse. *Journal of the Association for Information Science and Technology*, *70*(3), 296–300. DOI: 10.1002/asi.24116

Alperin, J. P., Schimanski, L. A., La, M., Niles, M. T., & McKiernan, E. C. (2020). The value of data and other non-traditional scholarly outputs in academic review, promotion, and tenure in Canada and the United States. In A. Berez-Kroeker, B. McDonnell, E. Koller, & L. Collister, *Open Handbook of Linguistic Data Management*. MIT Press. DOI: 10.17613/ye06-n045

Association of College and Research Libraries. (2015, February 9). *Framework for Information Literacy for Higher Education*. Association of College & Research Libraries. From http://www.ala.org/acrl/standards/ilframework

Atici, L., Kansa, S. W., Lev-Tov, J., & Kansa, E. C. (2013). Other People's Data: A Demonstration of the Imperative of Publishing Primary Data. Journal of Archaeological Method and Theory, 20(4), 663–681. DOI: 10.1007/s10816-012-9132-9

Barsky, E.. Davis, C., Darnell, A., Flynn, J., Goddard, L., Goodchild, M., Leahey, A., MacPherson, E., Roberge, P., Selman, B., Wilson, L. (2018). Recommendations for a National Dataverse Service. DOI: 10.14288/1.0385835

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, *23*(1), 4–17. DOI: 10.1016/j.intcom.2010.07.003

Baždarić, K., Vrkić, I., Arh, E., Mavrinac, M., Gligora Marković, M., Bilić-Zulle, L., Stojanovski, J., & Malički, M. (2021). Attitudes and practices of open data, preprinting, and peer-review—A cross sectional study on Croatian scientists. *PLOS ONE*, *16*(6), e0244529. DOI: 10.1371/journal.pone.0244529

Beissel-Durrant, G. (2004). *A Typology of Research Methods Within the Social Sciences* [Working Paper]. From http://eprints.ncrm.ac.uk/115/

Berman, E.A. (2017). An Exploratory Sequential Mixed Methods Approach to Understanding Researchers' Data Management Practices at UVM: Integrated Findings to Develop Research Data Services. *Journal of EScience Librarianship*, *6*(1), e1104. DOI: 10.7191/jeslib.2017.1104

Bishop, L. (2009). Ethical Sharing and Reuse of Qualitative Data. *Australian Journal of Social Issues*, *44*(3), 255–272. DOI: 10.1002/j.1839-4655.2009.tb00145.x

Bishop, L., & Kuula-Luumi, A. (2017). Revisiting Qualitative Data Reuse: A Decade On. *SAGE Open*, *7*(1), 2158244016685136. DOI: 10.1177/2158244016685136

Blankstein, M. (2022). *Ithaka S+R US Faculty Survey 2021*. Ithaka S+R. DOI: 10.18665/sr.316896

Borgman, C. L. (2010). The Digital Future is Now: A Call to Action for the Humanities. *Digital Humanities Quarterly*, *003*(4).

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. DOI: 10.1002/asi.22634

Borgman, C. L. (2015). *Big Data, Little Data, No Data: Scholarship in The Networked World.* The MIT Press.

Borgman, C. L. (2018). *Open Data, Grey Data, and Stewardship: Universities at the Privacy Frontier*. DOI: 10.15779/Z38B56D489

Borgman, C. L., & Bourne, P. E. (2022). Why It Takes a Village to Manage and Share Data. *Harvard Data Science Review*. DOI: 10.1162/99608f92.42eec111

Borgman, C. L., Sands, A. E., Darch, P. T., & Golshan, M. S. (2016). The durability and fragility of knowledge infrastructures: Lessons learned from astronomy: The Durability and Fragility of Knowledge Infrastructures: Lessons Learned from Astronomy. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–10. DOI: 10.1002/pra2.2016.14505301057

Borgman, C. L., Scharnhorst, A., & Golshan, M. S. (2019). Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *Journal of the Association for Information Science and Technology*, *70*(8), 888–904. DOI: 10.1002/asi.24172

Bow, C. (2019). Diverse socio-technical aspects of a digital archive of Aboriginal languages. *Archives and Manuscripts*, *47*(1), 94–112. DOI: 10.1080/01576895.2019.1570282

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. DOI: 10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2022). Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, *9*(1), 3–26. DOI: 10.1037/qup0000196

Brennan, T. (2017, October 15). *The Digital-Humanities Bust*. The Chronicle of Higher Education. From https://www.chronicle.com/article/the-digital-humanities-bust/

Briganti, J. S., Ogier, A., & Brown, A. M. (2020). Piloting a Community of Student Data Consultants that Supports and Enhances Research Data Services. *International Journal of Digital Curation*, 1–11. DOI: 10.2218/ijdc.v15i1.723

Briney, K., Goben, A., & Zilinski, L. (2015). Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies. *Journal of Librarianship and Scholarly Communication*, *3*(2), Article 2. DOI: 10.7710/2162-3309.1232

Buddenbohm, S., de Jong, M., Minel, J.-L., & Moranville, Y. (2021). Find research data repositories for the humanities—The data deposit recommendation service. *International Journal of Digital Humanities*, *1*(3), 343–362. DOI: 10.1007/s42803-021-00030-7

Bueno de la Fuente, Gema. (n.d.). *Challenges and strategies for the success of Open Science.* FOSTER: Facilitate Open Science Training for European Research. From https://www.fosteropenscience.eu/content/challenges-and-strategies-success-open-science

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T., & Schnapp, J. (2016). *Digital_humanities*. From https://mitpress.mit.edu/9780262528863/digital_humanities/

Callaghan, S. (2019). Research Data Publication: Moving Beyond the Metaphor. *Data Science Journal*, *18*(39). DOI: 10.5334/dsj-2019-039

Canadian Social Knowledge Institute. (2021, January 12). DHSI 2021 — Online Edition. *Digital Humanities Summer Institute*. From https://dhsi.org/dhsi-2021-online-edition-temp/

Carroll, S.R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. Sci Data 8 (108). DOI: 10.1038/s41597-021-00892-0

Carlson, J., Narlock, M., Blake, M., Herndon, J., & Imker, H. (2023). *The Art, Science, and Magic of the Data Curation Network: A Retrospective on Cross-Institutional Collaboration*. Maize Books. DOI: 10.3998/mpub.12782791

Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication*, *12*(2), 635–651. DOI: 10.1111/j.1083-6101.2007.00342.x

Case, M. M. (2008). Partners in Knowledge Creation: An Expanded Role for Research Libraries in the Digital Future. *Journal of Library Administration*, *48*(2), 141–156. DOI: 10.1080/01930820802231336

Cerda-Cosme, R., & Méndez, E. (2023). Analysis of shared research data in Spanish scientific papers about COVID-19: A first approach. *Journal of the Association for Information Science and Technology*, *74*(4), 402–414. DOI: 10.1002/asi.24716

Choi, S. Y., Lee, H., & Kang, Y. S. (2008). The effects of socio-technical enablers on knowledge sharing: An exploratory examination. *Journal of Information Science*, *34*(5), 742–754. DOI: 10.1177/0165551507087710

Choudhury, S., Huang, C., & Palmer, C. L. (1970). Updating the DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *15*(1), Article 1. DOI: 10.2218/ijdc.v15i1.721

Condon, P. B. (2015). Digital curation through the lens of disciplinarity: The development of an emerging field [Ph.D., Simmons College]. In *ProQuest Dissertations and Theses*. From http://search.proquest.com/docview/1662838077/abstract/567CF6864FD84294PQ/1

Conway, E., Giaretta, D., Lambert, S., & Matthews, B. (2011). Curating Scientific Research Data for the Long Term: A Preservation Analysis Method in Context. *International Journal of Digital Curation*, *6*(2), 38–52. DOI: 10.2218/ijdc.v6i2.204

Cooper, D., & Springer, R. (2019). *Data Communities: A New Model for Supporting STEM Data Sharing [Issue Brief]* (Copyright, Fair Use, Scholarly Communication, etc. 109.). From https://digitalcommons.unl.edu/scholcom/109

Cox, A. M., & Tam, W. W. T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, *70*(2), 142–157. DOI: 10.1108/AJIM-11-2017-0251

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038. DOI: 10.1098/rsta.2010.0165

Creswell, J. W. (2003). *Research design: Qualitative and quantitative approaches.* (2nd ed.). SAGE Publications.

Creswell, J. W. (2014). *Research design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed). SAGE Publications.

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. *PLOS ONE*, *12*(12), e0189288. DOI: 10.1371/journal.pone.0189288

Custers, B., & Bachlechner, D. (2017). Advancing the EU data economy: Conditions for realizing the full potential of data reuse. *Information Polity*, *22*(4), 291–309. DOI: 10.3233/IP-170419

Dalhousie University Libraries. (2019, November). *Institutional Research Data Management Strategy*. From https://dal.ca.libguides.com/c.php?g=257229&p=5167849

Davis, R. F., Gold, M. K., & Harris, K. D. (2020). *Curating Digital Pedagogy in the Humanities*. DOI: 10.17613/55A0-AM43

Devarakonda, R., Palanisamy, G., Green, J. M., & Wilson, B. E. (2011). Data sharing and retrieval using OAI-PMH. *Earth Science Informatics*, *4*(1), 1–5. DOI: 10.1007/s12145-010-0073-0

Devriendt, T., Shabani, M., & Borry, P. (2021). Data Sharing in Biomedical Sciences: A Systematic Review of Incentives. *Biopreservation and Biobanking*, *19*(3), 219–227. DOI: 10.1089/bio.2020.0037

Digital Curation Centre. (n.d.). *Overview of funders' data policies*. Retrieved January 12, 2022, from https://www.dcc.ac.uk/guidance/policy/overview-funders-data-policies

Digital Library Federation Forum 2017 Participants (n.d.). *The Santa Barbara Statement on Collections as Data*. Always Already Computational - Collections as Data. From https://collectionsasdata.github.io/statement/

Duke, C. S., & Porter, J. H. (2013). The Ethics of Data Sharing and Reuse in Biology. *BioScience*, *63*(6), 483–489. DOI: 10.1525/bio.2013.63.6.10

Duranti, L. (2001). *The Long-Term Preservation of Authentic Electronic Records*.

Edmiston, M., Coker, S., Jamilla, S., & Tshabalala, T. (2021, September 30). *The Pros and Cons of Open Data*. MERL Center. From https://merlcenter.org/guides/pros-and-cons-of-open-data/

Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, Ribes, D., Burton, M., & Calvert, S. (2013). *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Ann Arbor: Deep Blue. From http://hdl.handle.net/2027.42/97552

ERC Scientific Council. (2020, July 20). *ERC Scientific Council calls for open access plans to respect researchers' needs*. From https://erc.europa.eu/news/erc-scientific-council-calls-open-access-plans-respect-researchers-needs

Estill, L., Guiliano, J., Ortega, É., Terras, M., Verhoeven, D., & Layne-Worthey, G. (2022). The circus we deserve? A front row look at the organization of the annual academic conference for the Digital Humanities. *Digital Humanities Quarterly*, *016*(4). From http://www.digitalhumanities.org/dhq/vol/16/4/000643/000643.html

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416. DOI: 10.1002/asi.23480

Faniel, I.M. & Zimmerman, A. (2011). Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse, *International Journal of Digital Curation.* 6(1). DOI: 10.2218/ijdc.v6i1.172

Fecher, B., & Friesike, S. (2014). Open Science: One Term, Five Schools of Thought. In S. Bartling & S. Friesike (Eds.), *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing* (pp. 17–47). Springer International Publishing. DOI: 10.1007/978-3-319-00026-8_2

Federer, L. (2018). Defining data librarianship: A survey of competencies, skills, and training. *Journal of the Medical Library Association: JMLA*, *106*(3), 294–303. DOI: 10.5195/jmla.2018.306

Ferrari, T., Scardaci, D., & Andreozzi, S. (2018). The Open Science Commons for the European Research Area. In P.-P. Mathieu & C. Aubrecht (Eds.), *Earth Observation Open Science and Innovation* (pp. 43–67). Springer International Publishing. DOI: 10.1007/978-3-319-65633-5_3

First Nations Information Governance Centre. (2021, February 19). OCAP® and information governance. The First Nations Information Governance Centre. From https://fnigc.ca/what-we-do/ocap-and-information-governance

Flanders, J., & Hamlin, S. (2013). TAPAS: Building a TEI Publishing and Repository Service. *Journal of the Text Encoding Initiative*, *Issue 5*. DOI: 10.4000/jtei.788

Foote, K. D. (2020, November 25). A Brief History of Data Management. *DATAVERSITY*. From https://www.dataversity.net/brief-history-data-management/

Frank, R. D., Tyler, A. R. B., Gault, A., Suzuka, K., & Yakel, E. (1970). Issues of Privacy in Qualitative Video Data Reuse. *International Journal of Digital Curation*, *13*(1), 47–72. DOI: 10.2218/ijdc.v13i1.492

Garfield, E. (1980). Is Information Retrieval in the Arts and Humanities Inherently Different from That in Science? The Effect That ISI®'S Citation Index for the Arts and Humanities Is Expected to Have on Future Scholarship. *The Library Quarterly: Information, Community, Policy*, *50*(1), 40–57. JSTOR.

Garstki, K. (2022). Teaching for Data Reuse and Working toward Digital Literacy in Archaeology. *Advances in Archaeological Practice*, *10*(2), 177–186. DOI: 10.1017/aap.2022.3

Garwood, D. A., & Poole, A. H. (2019). Pedagogy and public-funded research: An exploratory study of skills in digital humanities projects. *Journal of Documentation*, *75*(3), 550–576. DOI: 10.1108/JD-06-2018-0094

Giglia, E. (2021). *Data stewardship, the Italian way*. DOI: 10.5281/ZENODO.5550099

GO FAIR Implementation Networks. (n.d.). *FAIR Principles*. GO FAIR. From https://www.go-fair.org/fair-principles/

Gravlee, C. C. (2022). Research Design and Methods in Medical Anthropology. In M. Singer, P. I. Erickson, & C. E. Abadía-Barrero (Eds.), *A Companion to Medical Anthropology* (1st ed., pp. 67–92). Wiley. DOI: 10.1002/9781119718963.ch4

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research. *Harvard Data Science Review*. DOI: 10.1162/99608f92.e38165eb

Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science*, *46*(4), 459–475. DOI: 10.1177/0165551519837182

Gregory, K., Ninkov, A., Ripp, C., Roblin, E., Peters, I. & Haustein, S. (2023). Tracing data: A survey investigating disciplinary differences in data citation (v2) [Preprint]. Zenodo. DOI: 10.5281/zenodo.7555265.

Griffin, T. M. (2020). Centering Graduate Students' Research Projects in Data Management Education: A Pilot Program. *Journal of Librarianship and Scholarly Communication*, *8*(1). DOI: 10.7710/2162-3309.2365

Groenewegen, D. (2016). *Research Data Ecosystem*. [Image]. DOI: 10.4225/03/575E42A4892B4

Grossetta Nardini, H. K., Batten, J., Funaro, M. C., Garcia-Milian, R., Nyhan, K., Spak, J. M., Wang, L., & Glover, J. G. (2019). Librarians as methodological peer reviewers for systematic reviews: Results of an online survey. *Research Integrity and Peer Review*, *4*(1), 23. DOI: 10.1186/s41073-019-0083-5

Habermann, T. (2023). Improving Domain Repository Connectivity. *Data Intelligence*, *5*(1), 6–26. DOI: 10.1162/dint_a_00120

Hajduk, G. K., Jamieson, N. E., Baker, B. L., Olesen, O. F., & Lang, T. (2019). It is not enough that we require data to be shared; we have to make sharing easy, feasible and accessible too! *BMJ Global Health*, *4*(4), e001550. DOI: 10.1136/bmjgh-2019-001550

Harvey, R. (2010). *Digital curation: A how to do it manual.* New York: Neal Schuman.

Hawkins, S. (Ed.). (2021). *Access and Control in Digital Humanities* (1st ed.). Routledge. DOI: 10.4324/9780429259616

Hedrick, T. E. (1988). Justifications for the sharing of social science data. *Law and Human Behavior*, *12*(2), 163–171. DOI: 10.1007/BF01073124

Heidorn, P. B. (2011). The Emerging Role of Libraries in Data Curation and E-science. *Journal of Library Administration*, *51*(7–8), 662–672. DOI: 10.1080/01930826.2011.601269

Hemphill, L., Pienta, A., Lafia, S., Akmon, D., & Bleckley, D. A. (2022). How do properties of data, their curation, and their funding relate to reuse? *Journal of the Association for Information Science and Technology*, *73*(10), 1432–1444. DOI: 10.1002/asi.24646

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *3*(1), 134–140. DOI: 10.2218/ijdc.v3i1.48

Hockey, S. (2016). Digital Humanities in the Age of the Internet: Reaching Out to Other Communities (chapter 5). In W. Mccarty & M. Deegan (Eds.), *Collaborative Research in the Digital Humanities* (0 ed.). Routledge. DOI: 10.4324/9781315572659

Hodges, P., Bonn, M., Sandler, M., & Wilkin, J. P. (2003). *Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of her Departure from the University of Michigan*. Michigan Publishing, University of Michigan Library. DOI: 10.3998/spobooks.bbv9812.0001.001

Hogan, M. (2015). The Archive as Dumpster. *Pivot: A Journal of Interdisciplinary Studies and Thought*, *4*(1). DOI: 10.25071/2369-7326.39565

Horik, R. V. (2019). *The Research Data Alliance and the Humanities*. Zenodo. DOI: 10.5281/ZENODO.3355145

Huchard, M., Laurent, A., Libourel, T., Madera, C., & Miralles, A. (2020). Exploiting Software Product Lines and Formal Concept Analysis for the Design of Data Lake Architectures. In A. Laurent, D. Laurent, & C. Madera (Eds.), *Data Lakes* (1st ed., pp. 41–56). Wiley. DOI: 10.1002/9781119720430.ch3

Huang C.K., Neylon C., Hosking R., Montgomery L., Wilson K.S., Ozaygen A., Brookes-Kenworthy C. Evaluating the impact of open access policies on research institutions. Elife. (2020, September 14). DOI: 10.7554/eLife.57067

Humanities Commons (n.d.). CORE Frequently Asked Questions. Humanities Commons. From https://hcommons.org/core/faq/

Humanities Commons. (n.d.). *Digital Pedagogy, Openness*. From https://digitalpedagogy.hcommons.org/introduction/conclusion-a-free-open-born-digital-collection/

Hwang, E., Kirkham, R., Marshall, K., Kharrufa, A., & Olivier, P. (2022). Sketching dialogue: Incorporating sketching in empathetic semi-Structured interviews for human-computer interaction research. *Behaviour & Information Technology*, 1–29. DOI: 10.1080/0144929X.2022.2113431

ICPSR. (n.d.). *Data Management & Curation*. Inter-University Consortium for Political and Social Research (ICPSR). From https://www.icpsr.umich.edu/web/pages/datamanagement/index.html

Imker, H. J., Luong, H., Mischo, W. H., Schlembach, M. C., & Wiley, C. (2021). An examination of data reuse practices within highly cited articles of faculty at a research university. *The Journal of Academic Librarianship*, *47*(4), 102369. DOI: 10.1016/j.acalib.2021.102369

Informatica Canada. (n.d.). *What is Data Stewardship | Informatica Canada*. Informatica Canada. From https://www.informatica.com/ca/resources/articles/what-is-data-stewardship.html

Internet Archive: Petabox. (n.d.). From https://archive.org/web/petabox.php

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLOS Biology*, *16*(3), e2005468. DOI: 10.1371/journal.pbio.2005468

Jacob, S. A., & Furgerson, S. P. (n.d.). *Writing Interview Protocols and Conducting Interviews: Tips for Students New to the Field of Qualitative Research*. 12.

Jeffery, R. B. (1998). Librarians as Generalists: Redefining Our Role in a New Paradigm. *Art Documentation: Journal of the Art Libraries Society of North America*, *17*(2), 25–29.

Jetten, M., Grootveld, M., Mordant, A., Jansen, M., Bloemers, M., Miedema, M., & Gelder, C. W. G. van. (2021). *Professionalising data stewardship in the Netherlands. Competences, training and education. Dutch roadmap towards national implementation of FAIR data stewardship*. Zenodo. DOI: 10.5281/zenodo.4486423

Kashyap, N. (2020). GitHub's Path to 128M Public Repositories With hard numbers. *Towards Data Science*. From https://towardsdatascience.com/githubs-path-to-128m-public-repositories-f6f656ab56b1

Kenney, A., & McGovern, N. Y. (2003). The Five Organizational Stages of Digital Preservation. In P. Hodges, M. Bonn, M. Sandler, & J. P. Wilkin (Eds.), *Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of her Departure from the University of Michigan*. Michigan Publishing, University of Michigan Library. DOI: 10.3998/spobooks.bbv9812.0001.001

Kervin, K., Cook, R. B., & Michener, W. K. (2014). The Backstage Work of Data Sharing. *Proceedings of the 18th International Conference on Supporting Group Work*, 152–156. DOI: 10.1145/2660398.2660406

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, *42*(5), 361–371. DOI: 10.1002/(SICI)1097-4571(199106)42:5<361::AID-ASI6>3.0.CO;2-#

Larivière V. & Sugimoto, C.R. (2018). Do authors comply when funders enforce open access to research? *Nature.* 562. 483–486. DOI: 10.1038/d41586-018-07101-w

Late, E., & Kekalainen, J. (2020). Use and users of a social science research data archive. *PLOS One*, *15*(8), e0233455. DOI: 10.1371/journal.pone.0233455

Lefebvre, A., Bakhtiari, B., & Spruit, M. (2020). Exploring research data management planning challenges in practice. *It-Information Technology*, *62*(1), 29–37. DOI: 10.1515/itit-2019-0029

Leigh, A., Makri, S., Taylor, A., Mulinder, A., & Hamdi, S. (2021). From Information to Knowledge Creation in the Archive: Observing Humanities Researchers' Information Activities. *Proceedings of the Association for Information Science and Technology*, *58*(1), 253–263. DOI: 10.1002/pra2.453

Lotan, E., Tschider, C., Sodickson, D. K., Caplan, A. L., Bruno, M., Zhang, B., & Lui, Y. W. (2020). Medical Imaging and Privacy in the Era of Artificial Intelligence: Myth, Fallacy, and the Future. *Journal of the American College of Radiology*, *17*(9), 1159–1162. DOI: 10.1016/j.jacr.2020.04.007

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). *Open Data Metrics: Lighting the Fire*. DOI: 10.5281/zenodo.3525349

Luff, R., Byatt, D., & Martin, D. (2015). *Review of the Typology of Research Methods within the Social Sciences* [Working Paper]. National Centre for Research Methods. From http://eprints.ncrm.ac.uk/3721/

Manganelli, M., Soldati, A., Martirano, L., & Ramakrishna, S. (2021). Strategies for Improving the Sustainability of Data Centers via Energy Mix, Energy Conservation, and Circular Energy. *Sustainability*, *13*(11), 6114. DOI: 10.3390/su13116114

Mannheimer, S., Newman, S., Coates, H. L., & Rinehart, A. (2021). Special Issue: 2020 Research Data Access and Preservation Summit. *Journal of EScience Librarianship*, *10*(1), 1197. DOI: 10.7191/jeslib.2021.1197

Milligan, I. (2022). *The Transformation of Historical Research in the Digital Age* (1st ed.). Cambridge University Press. DOI: 10.1017/9781009026055

Molloy, L. (2013, September 19). *UK Research Data Management: Overview to ADBU congress,* [Presentation]. From https://www.slideshare.net/lm_hatii/nn-molloy-adbumrdoverview20130919slideshare?from_action=save

Monserrate, S. G. (2022). The Cloud Is Material: On the Environmental Impacts of Computation and Data Storage. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, *Winter 2022*. DOI: 10.21428/2c646de5.031d4553

National Science Foundation. (2003). *"Knowledge lost in information", Report of the NSF Workshop on Research Directions for Digital Libraries* (NSF Award No. IIS-0331314).

Neubauer, B. E., Witkop, C. T., & Varpio, L. (2019). How phenomenology can help us learn from the experiences of others. *Perspectives on Medical Education*, *8*(2), 90–97. DOI: 10.1007/s40037-019-0509-2

Nitecki, D. A., & Davis, M. E. K. (2019). Expanding Academic Librarians' Roles in the Research Life Cycle. *Libri*, *69*(2), 117–125. DOI: 10.1515/libri-2018-0066

Nosek, B. (2019, June 11). Strategy for Culture Change. *Centre for Open Science: Strategy for Culture Change*. [Infographic]. From www.cos.io/blog/strategy-for-culture-changewww.cos.io/blog/strategy-for-culture-change

Nowak, R., & Haynes, J. (2018). Friendships with benefits? Examining the role of friendship in semi-structured interviews within music research. *International Journal of Social Research Methodology*, *21*(4), 425–438. DOI: 10.1080/13645579.2018.1431192

Nowviskie, B. (2011). *Evaluating Collaborative Digital Scholarship (or, Where Credit is Due) Journal of Digital Humanities*. From http://journalofdigitalhumanities.org/1-4/evaluating-collaborative-digital-scholarship-by-bethany-nowviskie/

Nowviskie, B. (2015). Digital Humanities in the Anthropocene. *Digital Scholarship in the Humanities*, *30*(suppl_1), i4–i15. DOI: 10.1093/llc/fqv015 https://doi.org/10.1093/llc/fqv015

OECD. (2020). *Enhanced Access to Publicly Funded Data for Science, Technology and Innovation*. OECD. DOI: 10.1787/947717bc-en

Office of the Chief Science Advisor of Canada. (2020). *ROADMAP FOR OPEN SCIENCE*. From https://science.gc.ca/site/science/en/office-chief-science-advisor/open-science/roadmap-open-science

Ohaji, I. K., Chawner, B., & Yoong, P. (2019). *The role of a data librarian in academic and research libraries*. University of Borås. From http://informationr.net/ir/24-4/paper844.html

Ontario Council of University Libraries Scholars Portal, Canadian Association of Research Libraries, Compute Canada, & Portage Network. (2019). *Repository Options in Canada: A Portage Guide*. Zenodo. DOI: 10.5281/ZENODO.3966349

Opdenakker, R. (2006). Advantages and Disadvantages of Four Interview Techniques in Qualitative Research. *Forum Qualitative Sozialforschung /* Forum: Qualitative Social Research, Vol 7, No 4 (2006): Qualitative Research in Ibero America. DOI: 10.17169/FQS-7.4.175https://doi.org/10.17169/FQS-7.4.175

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019). *Santa Barbara Statement on Collections as Data—Always Already Computational: Collections as Data*. DOI: 10.5281/zenodo.3066209https://doi.org/10.5281/zenodo.3066209

Padilla, T. G. (2018). Collections as data: Implications for enclosure. *College & Research Libraries News*, *79*(6), 296. DOI: 10.5860/crln.79.6.296https://doi.org/10.5860/crln.79.6.296

Padilla, T. (2016). Humanities Data in the Library: Integrity, Form, Access. *D-Lib Magazine*, *22*(3/4). DOI: 10.1045/march2016-padilla

Pandey, S. C., & Dutta, A. (2013). Role of knowledge infrastructure capabilities in knowledge management. *Journal of Knowledge Management*, *17*(3), 435–453. DOI: 10.1108/JKM-11-2012-0365

Pasquetto, I. (2018). From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses. [Dissertation]. From https://escholarship.org/uc/item/1sx7v77r

Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and Reuses of Scientific Data: The Data Creators' Advantage. *Harvard Data Science Review*, *1*(2). DOI: 10.1162/99608f92.fc14bf2d

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*(8). DOI: 10.5334/dsj-2017-008

Pasquetto, I. V., Sands, A. E., & Borgman, C. L. (2015). Exploring openness in data and science: What is "open," to whom, when, and why? *Proceedings of the Association for Information Science and Technology*, *52*(1), 1–2. DOI: 10.1002/pra2.2015.1450520100141

Pawlicka-Deger, U. (2021). Digital humanities and a new research culture: Between promoting and practicing open research data. In *Access and Control in Digital Humanities* (1st ed., pp. 40–57). Routledge. DOI: 10.4324/9780429259616

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS One*, *2*(3), e308. DOI: 10.1371/journal.pone.0000308

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. DOI: 10.7717/peerj.175

Piwowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Nature*, *473(7347)*, 285–285. DOI: 10.1038/473285a

Posner, M. (June 2015). "Humanities Data: A Necessary Contradiction," *Miriam Posner's Blog*, 25. From https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/

Poole, A., A. H. (2015). How has your science data grown? Digital curation and the human factor: a critical literature review. *Archival Science*, *15*(2), 101–139. DOI: 10.1007/s10502-014-9236-y

Poole, A. H. (2017). "A greatly unexplored area": Digital curation and innovation in digital humanities. *Journal of the Association for Information Science and Technology*, *68*(7), Article 7. DOI: 10.1002/asi.23743

Poole, A. H., & Garwood, D. A. (2018). "Natural allies": Librarians, archivists, and big data in international digital humanities project work. *Journal of Documentation*, *74*(4), 804–826. DOI: 10.1108/JD-10-2017-0137

Pronk, T. E. (2019). The Time Efficiency Gain in Sharing and Reuse of Research Data. *Data Science Journal*, *18*(10). DOI: 10.5334/dsj-2019-010

Purgar, M., Klanjscek, T., & Culina, A. (2022). Quantifying research waste in ecology. *Nature Ecology & Evolution*, *6*(9), 1390–1397. DOI: 10.1038/s41559-022-01820-0

Quan-Haase, A., & Martin, K. (2013). Digital curation and the networked audience of urban events: Expanding La Fiesta de Santo Tomás from the physical to the virtual environment. *International Communication Gazette*, *75*(5–6), 521–537. DOI: 10.1177/1748048513491910

Ramachandran, R., Bugbee, K., & Murphy, K. (2021). From Open Data to Open Science. *Earth and Space Science*, *8*(5). DOI: 10.1029/2020EA001562

Rebecca Grant. (2016, May 25). *RDA and the Digital Humanities*. From https://www.rd-alliance.org/rda-disciplines/rda-and-digital-humanities

Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public Data Archiving in Ecology and Evolution: How Well Are We Doing? *PLOS Biology*, *13*(11), e1002295. DOI: 10.1371/journal.pbio.1002295

Rowhani-Farid, A., Allen, M., & Barnett, A. G. (2017). What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review*, *2*(1), 4. DOI: 10.1186/s41073-017-0028-9

Ruediger, D., & MacDougall, R. (2023). *Are the Humanities Ready for Data Sharing?* Ithaka S+R. DOI: 10.18665/sr.318526

Safran, C. (2017). Update on Data Reuse in Health Care. *Yearbook of Medical Informatics*, *26*(01), 24–27. DOI: 10.15265/IY-2017-013

Salmi, H., Paju, P., Rantala, H., Nivala, A., Vesanto, A., & Ginter, F. (2021). The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective.

Historical Methods: A Journal of Quantitative and Interdisciplinary History, *54*(1), 14–28. DOI: 10.1080/01615440.2020.1803166

Savage, C. J., & Vickers, A. J. (2009). Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. *PLoS ONE*, *4*(9), Article 9. DOI: 10.1371/journal.pone.0007078

Scaramozzino, J. M., Ramírez, M. L., & McGaughey, K. J. (2012). A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University. *College & Research Libraries*, *73*(4), 349–365. DOI: 10.5860/crl-255

Schöch, C. (2014). *Big? Smart? Clean? Messy? Data In the Humanities*. DOI: 10.5281/ZENODO.8432

Schriml, L. M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., Hunter, C. I., Hurwitz, B. L., Kyrpides, N. C., Meyer, F., Mizrachi, I. K., Sansone, S.-A., Sutton, G., Tighe, S., & Walls, R. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Scientific Data*, *7*(1), 188. DOI: 10.1038/s41597-020-0524-5

Science Europe. (2020). Implementing Research Data Management Policies Across Europe: Experiences from Science Europe Member Organisations. DOI: 10.5281/zenodo.4915952

Science Staff. (2011). Challenges and Opportunities. *Science*, *331*(6018), 692–693. DOI: 10.1126/science.331.6018.692

Seastedt, K. P., Schwab, P., O'Brien, Z., Wakida, E., Herrera, K., Marcelo, P. G. F., Agha-Mir-Salim, L., Frigola, X. B., Ndulue, E. B., Marcelo, A., & Celi, L. A. (2022). Global

healthcare fairness: We should be sharing more, not less, data. *PLOS Digital Health*, *1*(10), e0000102. DOI: 10.1371/journal.pdig.0000102

Sengupta, U. (2022, November 1). *Towards a values-based data governance theory in the Social Economy in Ontario*. [Thesis]. TSpace. From https://hdl.handle.net/1807/124835

Sielemann, K., Hafner, A., & Pucker, B. (2020). The reuse of public datasets in the life sciences: Potential risks and rewards. *PeerJ*, *8*, e9954. DOI: 10.7717/peerj.9954

Smale, N. A., Unsworth, K., Denyer, G., Magatova, E., & Barr, D. (1970). A Review of the History, Advocacy and Efficacy of Data Management Plans. *International Journal of Digital Curation*, *15*(1), Article 1. DOI: 10.2218/ijdc.v15i1.525

Social Sciences and Humanities Research Council. (2012, November 29). *Research Data Archiving Policy*. From http://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx

Statistics Canada. (2020, September 23). *Data stewardship: An introduction*. https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062020013

Spiro, L. (2012). Opening Up Digital Humanities Education. In B. D. Hirsch (Ed.), Digital humanities pedagogy: practices, principles and politics (pp. 331–363). Cambridge, England: Open Book.

Stall, S., & Martone, M. (2020, February). *NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse: Workshop Summary*. https://datascience.nih.gov/data-ecosystem/nih-data-repository-workshop-summary

Stall, S., Martone, M. E., Chandramouliswaran, I., Crosas, M., Federer, L., Gautier, J., Hahnel, M., Larkin, J., Lowenberg, D., Pfeiffer, N., Sim, I., Smith, T., Van Gulick, A. E.,

Walker, E., Wood, J., Zaringhalam, M., & Zigoni, A. (2020). *Generalist Repository Comparison Chart*. DOI: 10.5281/ZENODO.3946720

Statistics Canada (Director). (2020, September 23). *Data stewardship: An introduction*. From https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062020013

Stigler, S. M. (2019). Data Have a Limited Shelf Life. *Harvard Data Science Review*. DOI: 10.1162/99608f92.f9a1e510

Stieglitz, S., Wilms, K., Mirbabaie, M., Hofeditz, L., Brenger, B., López, A., & Rehwald, S. (2020). When are researchers willing to share their data? – Impacts of values and uncertainty on open data in academia. *PLOS ONE*, *15*(7), e0234172. DOI: 10.1371/journal.pone.0234172

Strupler, N., & Wilkinson, T. C. (2017). Reproducibility in the Field: Transparency, Version Control and Collaboration on the Project Panormos Survey. *Open Archaeology*, *3*(1). DOI: 10.1515/opar-2017-0019

Sturgeon, D. (2018). Unsupervised identification of text reuse in early Chinese literature. *Digital Scholarship in the Humanities*, *33*(3), 670–684. DOI: 10.1093/llc/fqx024

Sukumara, P. T., & Metoyer, R. (2019). Replication and Transparency of Qual- itative Research from a Constructivist Perspective. *OSF Preprints*. From osf.io/6efvphttps://osf.io/6efvp

Sunikka, A. (2019). Organising RDM and Open Science Services. *International Journal of Digital Curation*, *14*(1), 180–193. DOI: 10.2218/ijdc.v14i1.641

Tayler, F., Michell, M., Ripp, C., & Dangoisse, P. (2022). *Data Primer: Making Digital Humanities Research Data Public // Manuel d'introduction aux données : Rendre*

publiques les données de recherche en sciences humaines numériques [dataset]. Borealis. DOI: 10.5683/SP3/OMLXTZ

Teherani, A., Martimianakis, T., Stenfors-Hayes, T., Wadhwa, A., & Varpio, L. (2015). Choosing a Qualitative Research Approach. *Journal of Graduate Medical Education*, *7*(4), 669–670. DOI: 10.4300/JGME-D-15-00414.1

Tempini, N. (2017). Till data do us part: Understanding data-based value creation in data-intensive infrastructures. *Information and Organization*, *27*(4), 191–210. DOI: 10.1016/j.infoandorg.2017.08.001

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLOS ONE*, *10*(8), e0134826. DOI: 10.1371/journal.pone.0134826

Tenopir, C., Hughes, D., Allard, S., Frame, M., Birch, B., Baird, L., Sandusky, R., Langseth, M., & Lundeen, A. (2015). Research Data Services in Academic Libraries: Data Intensive Roles for the Future? *Journal of EScience Librarianship*, *4*(2). DOI: 10.7191/jeslib.2015.1085

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE*, *15*(3), e0229003. DOI: 10.1371/journal.pone.0229003

Tenopir, C., Sandusky, R. J., Allard, S., & Birch, B. (2014). Research data management services in academic research libraries and perceptions of librarians. *Library & Information Science Research*, *36*(2), 84–90. DOI: 10.1016/j.lisr.2013.11.003

Tewell, E., Mullins, K., Tomlin, N., & Dent, V. (2017). Learning about Student Research Practices through an Ethnographic Investigation: Insights into Contact with Librarians and Use of Library Space. *Evidence Based Library and Information Practice*, *12*(4), 78–101. DOI: 10.18438/B8MW9Q

Tollefson, J. & Van Noorden. R. (2022) "US Government Reveals Big Changes to Open-Access Policy." *Nature*, 609(7926) pp. 234–35. DOI: 10.1038/d41586-022-02351-1

The Digital Curation Centre (DCC). (n.d.). *Curation Lifecycle Model*. The Digital Curation Centre. https://www.dcc.ac.uk/guidance/curation-lifecycle-model

The ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. DOI: 10.1038/nature05874

The White House. (n.d.). *FACT SHEET: Biden-Harris Administration Announces New Actions to Advance Open and Equitable Research* [Press release]. https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research/

Turp, C., Wilson, L., Pascoe, J., & Garnett, A. (2020). The Fast and the FRDR: Improving Metadata for Data Discovery in Canada. *Publications*, *8*(2), 25. DOI: 10.3390/publications8020025

Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., Hayashi, K., O'Donnell, D., Digital Science, & Hook, D. (2016). *The State of Open Data Report*. Digital Science. DOI: 10.6084/M9.FIGSHARE.4036398

Treasury Board of Canada Secretariat. (2011, March 17). *Minister Day Launches Open Data Portal*. https://web.archive.org/web/20110706181845/

TU Delft Library. (n.d.). Data stewardship. TU Delft. From https://www.tudelft.nl/library/research-data-management/r/support/data-stewardship

University of British Columbia. (2019). Documentation: Open UBC/OpenScholarship. In *UBC Wiki*. From wiki.ubc.ca/Documentation:Open_UBC/OpenScholarship

University of Helsinki/ Lindholm, T., Ojanen, M., & Siipilehto, L. (2020, September 3). [Infographic]. What is research data management (RDM)? *THINK OPEN.* From blogs.helsinki.fi/thinkopen/know-your-data-rdm-series-1/

Van de Sandt, S., Dallmeier-Tiessen, S., Lavasa, A., & Petras, V. (2019). The Definition of Reuse. *Data Science Journal*, *18*(22). DOI: 10.5334/dsj-2019-022

Velkova, J. (2019). Data Centres as Impermanent Infrastructures. *Culture Machine*, *18*(11). From http://culturemachine.net/vol-18-the-nature-of-data-centers/data-centers-as-impermanent/

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, *88*, 428–436. DOI: 10.1016/j.jbusres.2017.12.043

Vision, T. J. (2010). Open Data and the Social Contract of Scientific Publishing. *BioScience*, *60*(5), 330–331. DOI: 10.1525/bio.2010.60.5.2

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, *8*(7), e67332. DOI: 10.1371/journal.pone.0067332

Walsh, K. (2012). Board Editorial: Quantitative vs qualitative research: A false dichotomy. *Journal of Research in Nursing*, *17*(1), 9–11. DOI: 10.1177/1744987111432053

Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. *Journal of the Association for Information Science and Technology*, *72*(9), 1161–1182. DOI: 10.1002/asi.24483

Wang, Y., & Mi, J. (2012). Searchability and Discoverability of Library Resources: Federated Search and Beyond. *College & Undergraduate Libraries*, *19*(2–4), 229–245. DOI: 10.1080/10691316.2012.698944

Ward, C., Freiman, L., Jones, S., Molloy, L., & Snow, K. (2011). Making Sense: Talking Data Management with Researchers. *International Journal of Digital Curation*, *6*(2), 265–273. DOI: 10.2218/ijdc.v6i2.202

Watson-Boone, R. (1994). The information needs and habits of humanities scholars. *RQ*, *34*(2), 203–216.

Weingart, S. B., & Eichmann-Kalwara, N. (2017). What's Under the Big Tent?: A Study of ADHO Conference Abstracts. *Digital Studies/Le Champ Numérique*, *7*(1), 6. DOI: 10.16995/dscn.284

Whitelaw, M. (2018). Mashups and Matters of Concern: Generative Approaches to Digital Collections. *Open Library of Humanities*, *4*(1), Article 1. DOI: 10.16995/olh.291

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The Poor Availability of Psychological Research Data for Reanalysis. *American Psychologist*, *61*(7), 726–728. DOI: 10.1037/0003-066X.61.7.726

Wildemuth, B. M. (2017). *Applications of Social Research Methods to Questions in Information and Library Science* (2nd Edition).

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. 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*(1), 160018. DOI: 10.1038/sdata.2016.18

Williams, C. (2011). Research Methods. *Journal of Business & Economics Research (JBER)*, *5*(3). DOI: 10.19030/jber.v5i3.2532

Willinsky, J., & Wolfson, L. (2001). The Indexing of Scholarly Journals: A Tipping Point for Publishing Reform? *The Journal of Electronic Publishing*, *7*(2). DOI: 10.3998/3336451.0007.202

Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences: Red Flags in Data: Learning from Failed Data Reuse Experiences. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–6. DOI: 10.1002/pra2.2016.14505301126

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. *Library & Information Science Research*, *39*(3), 224–233. DOI: 10.1016/j.lisr.2017.07.008

Zakarya, M. (2018). Energy, performance and cost-efficient datacenters: A survey. *Renewable and Sustainable Energy Reviews*, *94*, 363–385. DOI: 10.1016/j.rser.2018.06.005

Conflict of Interest Statement

The author of this thesis has been employed by a Canadian DRI, the Digital Research Alliance of Canada, since 2022.

APPENDIX A

ADHO CONFERENCE DATA 2018, 2019, 2020

Authors listed on the website were listed by type of presentation. My search strategy was to use those parameters to narrow down a list of presenters to exclude poster presenters as they may not necessarily be considered DH researchers. The inclusion criteria for the search in each conference year included the terms “Invited Institute Lecturer”, “Invited speaker”, “DHSI Conference and Colloquium Speaker”, “paper”, “short paper” or “presentation”. DH2020 did not take place due to COVID-19 restrictions, but speakers were accepted for the conference that had been planned to take place at Carleton University and the University of Ottawa in Ottawa, Canada. The sample of names was sourced and imported as textual data from https://dh2020.adho.org/ and saved into a spreadsheet. DH2019 took place at Universiteit Utrecht, in Utrecht, Netherlands the sample of names was sourced and imported as textual data from https://staticweb.hum.uu.nl/dh2019/dh2019.adho.org/index.html and saved into a spreadsheet. DH2018 took place at El Colegio de México, Universidad Nacional Autónoma de México, Mexico City, Mexico, and the sample of names was sourced and imported as textual data from https://dh2018.adho.org/en/ and saved into a spreadsheet.

APPENDIX B

PARTICIPANT RECRUITMENT EMAIL

Subject line:

University of Ottawa study - Data Reuse Among Digital Humanities Scholars

Dr. [Last name],

My name is Lina Harper, and I am a master's student working under the supervision of Dr. Stefanie Haustein at the School of Information Studies at the University of Ottawa. I am conducting a study on data reuse among digital humanities scholars (DH) – and would like to invite you to become a participant. Your name was randomly selected because you presented at one or more of the last three years of the Alliance of Digital Humanities Organizations conferences. If you **consider yourself a DH researcher**, I would like to invite you to an interview on the use and/or reuse of data in the context of your DH research in the next weeks. Your participation will consist of a single one-on-one semi-structured interview that should take no more than 50 minutes of your time. The interview will be conducted using Microsoft Team and will be audio-recorded to ensure accurate transcription and analysis. Please note that I have received ethics clearance through the University of Ottawa Research Ethics Board (file # S-04-21-6356). If you would like to be a participant in this study, please reply to this email with a suggested date and time (9 am - 7 pm EST) to meet virtually. I will follow up with meeting materials including the interview questions and a meeting link. If you require additional information to assist you in reaching a decision about participation, please do not hesitate to reply to this email.

APPENDIX C

LETTER OF INFORMED CONSENT

Thank you for agreeing to participate in the “Data reuse among digital humanities scholars” study. The purpose of this study is to identify motivations and barriers in data reuse among digital humanities scholars. The project aims to enhance current thinking and provide insight for information and library science practitioners on the praxis of digital humanities and data curation. More widely, it is hoped that the collected qualitative data will be useful to inform scholarly communications. The study will conclude by reporting the results in the form of pragmatic approaches to data curation based on empirical evidence. The final analysis may result in recommendations for building data curation capacity in the digital humanities.

Use of data: The audience for this thesis paper will be researchers and makers in the digital humanities, librarians, funders, and data repository managers. The study will be conducted in the context of a Master of Arts in Information Studies thesis by Lina Marie Harper, under the supervision of Dr. Stefanie Haustein. Your participation consists of taking part in one (1) individual video-conferencing interview. The 45-60 minutes interview will be audio-recorded through the video conferencing application. It will focus on your perceptions and use of data, and on data reuse.

Risks: There are no foreseeable risks associated with participation in this study. All personal information, including your name, affiliations, audio recordings of interviews, transcribed interviews, and notes will remain strictly confidential. Anonymity will be assured using an alias unique to you. Audio recordings of interviews, notes,

coding, and annotations will be stored on a home office desktop computer and any printed backups will be stored in a locked file folder in the supervisor's office. Both physical and digital data will be kept for five (5) years after which they will be destroyed. Results of this study may be published in academic journals and presented at conferences. Any quotations will be attributed to an alias to ensure protect participants anonymity.

Withdrawal of participation: You may withdraw at any time.

If you have any questions about the study, please contact Lina Marie Harper (PI) or her supervisor, Dr. Haustein.

Informed consent agreement

"I, _____(full name), agree to participate in the research conducted by Lina Marie Harper (Principal Investigator), and supervised by Dr. Stefanie Haustein.

I understand that the contents related to my participation in this study will only be used by the PI and the supervisor. My participation in this study is voluntary and I am free to withdraw from the study at any time or refuse to answer any questions without consequence. If I choose to withdraw from the study, the data collected from my interview until the time of withdrawal will be destroyed and will not be used."

Please place an X next to any of the following statements to confirm your consent:

- I consent to being audio-recorded (if not in agreement, interview will not take place).

- I acknowledge that I am receiving no compensation for participating in this study.
- I consent to being quoted in the final report.

Please type out your name, date it, and draw or append a digitized signature in the spaces below:

Participant name: _____ Date: _____

Signature: _____

APPENDIX D

INTERVIEW SCHEDULE

SCRIPT

1. Context: brief intro about research topic
2. “Thank you for signing the consent form.” Consent form highlights: “Participation in this interview for this research project is voluntary and you may stop at any time or decline to answer any question including recordings will be stored are included in the consent form for your information, as are the contact details for reaching me or my thesis supervisor.
3. “I will now press record and we will begin the interview.”

BLOCK I: DATA USE AND REUSE IN RESEARCH

1. Are you a digital humanist or do you “do” digital humanities in your work?
2. Have you ever reused data collected by others? What kind of data did you reuse? (Never reused data, interviewer to move to interview schedule a)
3. Have you ever searched for data collected by others?
4. Have you ever encountered challenges in reusing data collected by others? (if yes, ask to expand).
5. Please briefly describe your main research method: quantitative, qualitative, mixed method?
6. What type of data do you generally use? (prompt: examples of data included interview transcripts, datasets including structure or unstructured, numerical, or textual data).

7. Earlier in the interview, I asked you if you had ever reused data collected by others.
    a. Please describe the data you reused.
    b. How did you locate the data that you wanted to reuse?

(prompt: if found by searching online, serendipitously or via the literature)

    c. (if not covered in the answers to the above) Where did you look for and/or find data to reuse? (prompt: examples may include social networks, websites, or databases)
    d. Describe your experience in terms of accessing the dataset(s).

(Prompt) Was is easy to find it? Access it? Sort through and understand it? Was it described well, was there metadata?)

    e. How reliable or trustworthy was the data?

(prompt) If the desired dataset was not used in the end, what was the main reason for not using the dataset for your research project?

Do you want to add anything regarding how you use data in research?

BLOCK II: RESEARCH DATA MANAGEMENT

8. Do you formalize the way in which you collect, analyse, store, and manage data, for example in a Data Management Plan? (prompt: a DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed).

(If yes)

   a. Did you get help in writing the DMP, for example, help from a librarian, archivist, student, or another team member? Did you base your DMP on a template or an exemplar, some kind of automated DMP maker?

9. Looking back at a completed research project, can you share some of the lessons learned in managing research data?
10. Have you ever taken a class, attended a workshop or other training in research data management?

    (if no)

    a. Would you like to know more about data curation or research data management?

    OR

    b. What kind of training would you prefer – for example online course, in person course? Formal or informal workshops? Watching or listening to a recording of someone presenting on the topic?

BLOCK III: DATA SHARING AND REUSE

11. Have you ever shared, or are you planning to share, your data?

    (if yes)

    c. Where did you share it (for example, a university repository, a non-profit organization's repository, an online academic social network, an online data, or technology community)?

d. Do you know if others reused data that you collected? What was it? How did you find out about it? Did you read the research produced from the reused?

(if no)

a. What would it take for you to share your data?

BLOCK IV: CLOSING

12. Do you have anything to add concerning DH and data reuse – do you have opinions on data and dh and what it could look like?
13. Are there any questions that you wanted to get back to or is there anything else that you'd like to talk about? Do you have any questions for me?

Thank you for your time. I will now stop recording.