Data Migration to Open Journal Systems (OJS) using R

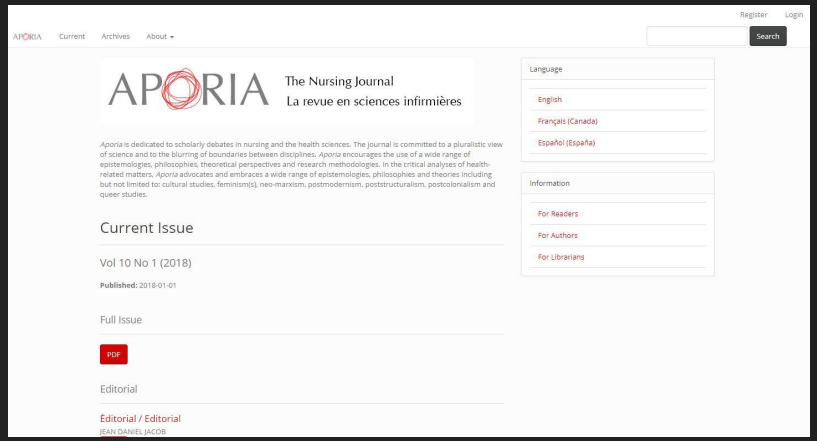
Yoo Young Lee
Web and Digital Initiatives Librarian
University of Ottawa

Access 2018

Project Background

- Aporia: peer-reviewed and open access journal in nursing and health sciences
- First journal that uOttawa Library hosted
- In-house legacy publishing system: stand-alone web server
  - JavaServer Pages (JSP)
  - Java Classes
  - Apache Tomcat

Project Background



- Aporia is available on the uOttawa Scholars Portal OJS
- Once the latest issue is published, Scholarly Communications Librarian will officially announce and update its new location

Project Goal

- To figure out and convert metadata in HTML into XML

HTML

```
<div id="abstract_page">
  <div id="content">
        <p class="pad_top60 titre">TITLE</p>
        <p class="auteur">AUTHOR</p>
        <p class="pad_top20 bold">Abstract</p>
            <p>ABSTRACT</p>
        <p class="pad_top20 bold">Key Words</p>
            <p>KEYWORDS</p>
        <p class="pad_top20">
            <a href=".../articles/YEAR_MONTH/..."></a>
        </p>
  </div>
</div>
```

XML

```
<article>
  <title>TITLE</title>
  <abstract>ABSTRACT</abstract>
  <keywords>
    <keyword>KEYWORD1</keyword>
    <keyword>KEYWORD2</keyword>
  </keywords>
  <authors>
    <author>
      <givenname>FIRST NAME</givenname>
      <familyname>LAST NAME</familyname>
    </author>
  </authors>
</article>
```

Project Tool & Packages

- Rcrawler: “R package for web crawling websites and extracting structured data which can be used for a wide range of useful applications, like web mining and text mining”
- data.table: “Extension of data.frame to aggregate large data”
- dplyr: “Grammar of data manipulation”
- XML: “Reading and creating XML documents in R”
- stringr: “Data cleaning and preparation tasks for strings”

Challenge 1: Crawl web data

HTML

```
<div id="abstract_page">
  <div id="content">
        <p class="pad_top60 titre">TITLE</p>
        <p class="auteur">AUTHOR</p>
        <p class="pad_top20 bold">Abstract</p>
            <p>ABSTRACT</p>
        <p class="pad_top20 bold">Key Words</p>
            <p>KEYWORDS</p>
        <p class="pad_top20">
            <a href=".../articles/YEAR_MONTH/..."></a>
        </p>
  </div>
</div>
```

Rcrawler

- **Rcrawler()**: Parse an entire website with all web pages and extract all data with a single command line

- Elements per pattern
  - CSS selectors
  - XPath

Challenge 1: Crawl web data

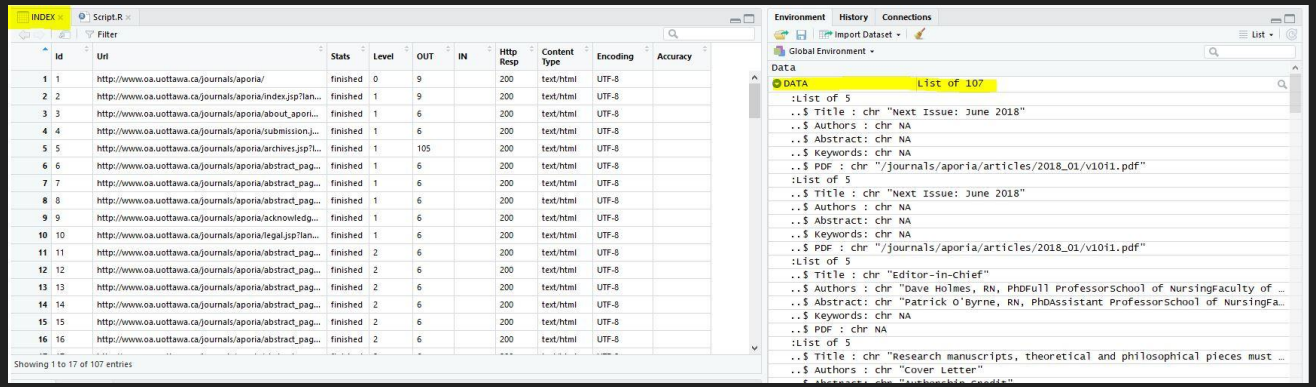
Command:

Rcrawler(**website**="http://www.oa.uottawa.ca/journals/aporia/", **ExtractXpathPat** = c("//p[1]", "//p[2]", "//p[4]", "//p[6]", "//a[contains(@href, 'articles')]/@href"), **PatternsNames** = c("title", "authors", "abstract", "keywords", "url"))

HTML

```
<div id="abstract_page">
  <div id="content">
        1 <p class="pad_top60 titre">TITLE</p>
        2 <p class="auteur">AUTHOR</p>
          <p class="pad_top20 bold">Abstract</p>
            4 <p>ABSTRACT</p>
          <p class="pad_top20 bold">Key Words</p>
            6 <p>KEYWORDS</p>
          <p class="pad_top20">
              <a href=".../articles//YEAR_MONTH/..."></a>
           </p>
  </div>
</div>
```

Challenge 1: Crawl web data



Outcome:

- INDEX variable: Data frame representing the general URL index with their details like content type, HTTP state, # of out- and in-links, encoding type, and level)
- **DATA variable**: Lists of extracted contents
- Directory: All crawled web pages with same numeric “id” in INDEX

Challenge 2: Convert list to data frame

DATA variable

- List of 5
  - title
  - abstract
  - authors
  - keywords
  - url

data.table

- Package to facilitate fast aggregation of large data
- **rbindlist()**: Make one data table from a list of many

Terms:

- List: It contains elements of different types like title, abstract, authors, keywords, url as one list
- Data Frame: Most common way to store data in R and list of equal-length vectors (rows & columns)

Challenge 2: Convert list to data frame

Command:

original_data_frame <- rbindlist(DATA)

Outcome:

- Data Frame named **original_data_frame**

	Title	Authors	Abstract	Ke
1	Next Issue: June 2018	NA	NA	NA
2	Next Issue: June 2018	NA	NA	NA
3	Editor-in-Chief	Dave Holmes, RN, PhDFull ProfessorSchool of NursingFa...	Patrick O'Byrne, RN, PhDAssistant ProfessorSchool of Nu...	NA
4	Research manuscripts, theoretical and philosophical pie...	Cover Letter	Authorship Credit	Co
5	Aporia: The Nursing Journal is available through the foll...	Volume 9	Volume 7	Vo
6	Les représentations du VIH, des personnes vivant avec le...	Stephany Cator & Marilou Gagnon	Depuis la découverte du virus, les manuels en soins infir...	ar
7	Accueillir et énoncer la mort aux soins intensifs : jeux d'a...	Valérie Drolet, Nicolas Vonarx & Diane Tapp	Cet article propose de réfléchir sur la place, l'énonciatio...	Ce
8	Speak Medicine Inc.! : Enjeux éthiques et politiques du t...	Geneviève McCready & Lucie-Catherine Ouimet	Le recours aux infirmières praticiennes spécialisées est p...	Bo
9	Launching a new scientific journal has required extensiv...	Jasmine Bouchard, Web Initiatives LibrarianSean Boxer, ...	NA	NA
10	Authors who publish with APORIA agree to the followin...	Authors retain copyright and grant the journal right of f...	NA	NA
11	Philosophie, méthode dialectique et théorie critique : de...	Pawel J. Krol & Sophie Boisvert	À l'instar de l'investigation scientifique, la philosophie n...	di
12	Vers une meilleure compréhension des dimensions médi...	Pierre Pariseau-Legault	Plusieurs balises juridiques agissent en périphérie de la ...	ac
13	Critical bioethics in the time of epidemic: The case of the...	Jennifer M. Kilty, Michael Orsini & Péter Balogh	This article highlights the ethically uncertain and emotio...	cri
14	De la théorie postcoloniale en sciences infirmières : une ...	Marie-Pier Labelle & Patrick Martin	Fortement influencée par la théorie poststructuraliste et...	én
15	Exploring the Potential Contribution of Actor-Network T...	Annie Rioux-Dubois & Amélie Perron	Nurse Practitioners (NPs) are clinically effective and safe. ...	ac
16	An Emancipating-salutogenesis conceptual framework ...	Margareth Santos Zanchetta, Melissa Stevenson, Vera N...	This paper presents a conceptual framework and model ...	Ar

Challenge 3: Data cleaning

Command:

```
# Keep completed rows
modified_data_frame <- original_data_frame[complete.cases(original_data_frame), ]
```

	Title	Authors	Abstract	Ke
1	Next Issue: June 2018	NA	NA	Na
2	Next Issue: June 2018	NA	NA	Na
3	Editor-in-Chief	Dave Holmes, RN, PhDFull ProfessorSchool of NursingFa...	Patrick O'Byrne, RN, PhDAssistant ProfessorSchool of Nu...	Na
4	Research manuscripts, theoretical and philosophical pie...	Cover Letter	Authorship Credit	Co
5	Aporia: The Nursing Journal is available through the foll...	Volume 9	Volume 7	Vo
6	Les représentations du VIH, des personnes vivant avec le...	Stephany Cator & Marilou Gagnon	Depuis la découverte du virus, les manuels en soins infir...	ar
7	Accueillir et énoncer la mort aux soins intensifs : jeux d'a...	Valérie Drolet, Nicolas Vonarx & Diane Tapp	Cet article propose de réfléchir sur la place, l'énonciatio...	Ge
8	Speak Medicine Inc.! : Enjeux éthiques et politiques du t...	Geneviève McCready & Lucie-Catherine Ouimet	Le recours aux infirmières praticiennes spécialisées est p...	Bo
9	Launching a new scientific journal has required extensiv...	Jasmine Bouchard, Web Initiatives LibrarianSean Boxer, ...	NA	Na
10	Authors who publish with APORIA agree to the followin...	Authors retain copyright and grant the journal right of f...	NA	Na
11	Philosophie, méthode dialectique et théorie critique : de...	Pawel J. Krol & Sophie Boisvert	À l'instar de l'investigation scientifique, la philosophie n...	di
12	Vers une meilleure compréhension des dimensions médi...	Pierre Pariseau-Legault	Plusieurs balises juridiques agissent en périphérie de la ...	ac
13	Critical bioethics in the time of epidemic: The case of the...	Jennifer M. Kilty, Michael Orsini & Péter Balogh	This article highlights the ethically uncertain and emotio...	cri
14	De la théorie postcoloniale en sciences infirmières : une ...	Marie-Pier Labelle & Patrick Martin	Fortement influencée par la théorie poststructuraliste et...	én
15	Exploring the Potential Contribution of Actor-Network T...	Annie Rioux-Dubois & Amélie Perron	Nurse Practitioners (NPs) are clinically effective and safe. ...	ac
16	An Emancipating-salutogenesis conceptual framework ...	Margareth Santos Zanchetta, Melissa Stevenson, Vera N...	This paper presents a conceptual framework and model ...	Ar

	Title	Authors	Abstract	Ke
1	Les représentations du VIH, des personnes vivant avec le...	Stephany Cator & Marilou Gagnon	Depuis la découverte du virus, les manuels en soins infir...	ar
2	Accueillir et énoncer la mort aux soins intensifs : jeux d'a...	Valérie Drolet, Nicolas Vonarx & Diane Tapp	Cet article propose de réfléchir sur la place, l'énonciatio...	Ge
3	Speak Medicine Inc.! : Enjeux éthiques et politiques du t...	Geneviève McCready & Lucie-Catherine Ouimet	Le recours aux infirmières praticiennes spécialisées est p...	Bo
4	Philosophie, méthode dialectique et théorie critique : de...	Pawel J. Krol & Sophie Boisvert	À l'instar de l'investigation scientifique, la philosophie n...	di
5	Vers une meilleure compréhension des dimensions médi...	Pierre Pariseau-Legault	Plusieurs balises juridiques agissent en périphérie de la ...	ac
6	Critical bioethics in the time of epidemic: The case of the...	Jennifer M. Kilty, Michael Orsini & Péter Balogh	This article highlights the ethically uncertain and emotio...	cri
7	De la théorie postcoloniale en sciences infirmières : une ...	Marie-Pier Labelle & Patrick Martin	Fortement influencée par la théorie poststructuraliste et...	én
8	Exploring the Potential Contribution of Actor-Network T...	Annie Rioux-Dubois & Amélie Perron	Nurse Practitioners (NPs) are clinically effective and safe. ...	ac
9	An Emancipating-salutogenesis conceptual framework ...	Margareth Santos Zanchetta, Melissa Stevenson, Vera N...	This paper presents a conceptual framework and model ...	Ar
10	A Critical Analysis of the Use of Remote Presence Robots...	Louise Racine	The exponential proliferation of e-learning programs ha...	di
11	"So far it's been choosing which side effects I want or I c...	Marilou Gagnon & Dave Holmes	Despite the availability of new antiretroviral drugs and t...	ac
12	The utility of queer theory in reconceptualising ageing i...	Maurice Nagington	With the advent of effective anti-viral regimes ageing wi...	Ag
13	De l'économie de la santé publique à la bioéconomie: a...	Pierre-Marie David	Les savoirs d'économie de la santé et de santé publique ...	Bi
14	Is Queer Sex Education in Ontario Finally Out of the Clo...	Cameron McKenzie	In 2010, Ontario's Ministry of Education introduced a rev...	Ou
15	Gouvernementalité de la « bonne » mère contemporaine...	Marguerite Soulière & Denise Moreau	Ce texte propose une analyse qui conjugue le versant p...	«
16	Une approche collaborative interprofessionnelle pour ré...	MONIQUE BENOIT, ANNE MARISE LAVOIE, & JEAN DRAG...	La plupart des personnes âgées dépendent du personn...	ab

Outcome:

- Modified Data Frame named **modified_data_frame**

Challenge 4: Data Modification

title	authors	abstract	keywords	url

dplyr

- **mutate()**: adds new variables that are functions of existing variables
- **case_when()**: vectorize multiple if and else if statements

Core function

- **grepl()**: pattern matching

Challenge 4: Data Modification

Command:

```
# Add Volume, Number, Year, and Published Date
# Create a new column called volume with an empty value
modified_data_frame %>% mutate(volume=case_when(
        grepl(“/2018_01/”, url) ~ “10”,
        grepl(“/2017_07/”, url) ~ “9”,
        ...
))

modified_data_frame %>% mutate(number=case_when(
        grepl(“/2018_01/”, url) ~ “1”,
        grepl(“/2017_07/”, url) ~ “2”,
        ...
))
```

dplyr

- **mutate()**: adds new variables that are functions of existing variables
- **case_when()**: vectorize multiple if and else if statements

Core function

- **grepl()**: pattern matching

Challenge 4: Data Modification

Command:

final_data_frame <- modifed_data_frame[,c(1,3,4,2,6,7,8,9)]

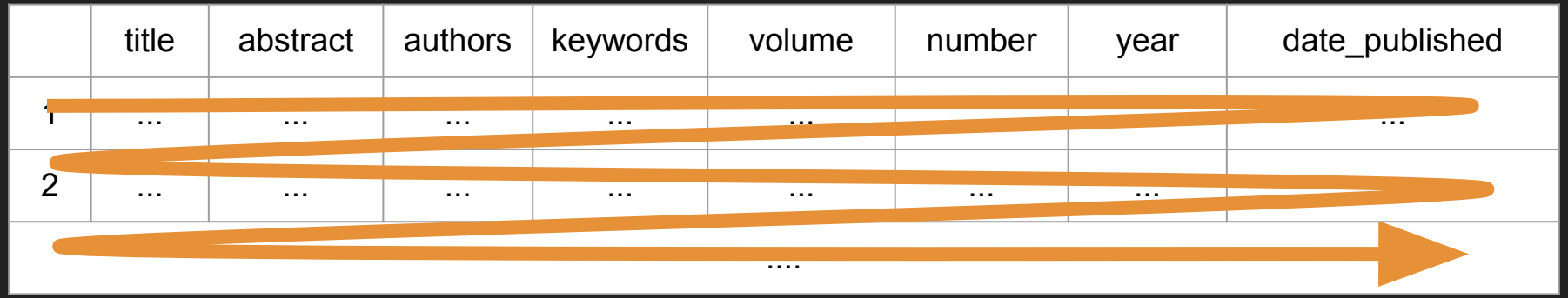
1	2	3	4	5	6	7	8	9
title	authors	abstract	keywords	url	volume	number	year	date_published

title	abstract	keywords	authors	volume	number	year	date_published

Challenge 5: Data frame to XML

Concept:

- Develop a custom function to transform metadata from data frame to XML applying functions in other packages
  - Iterate by rows => **lapply()**
  - Once in a row, check by column => **for loops**



Challenge 5: Data frame to XML

Command:

```
aporia_xml <- function(data, file=NULL) {
        aporia = xmlTree("articles", namespaces = list("http://pkp.sfu.ca",
                                                       xsi="http://www.w3.org/2001/XMLSchema-instance"),
                                 attrs=c("xsi:schemaLocation"="http://pkp.sfu.ca native.xsd"))


        ....
}
```

Outcome:

```
<?xml version="1.0" encoding="UTF-8"?>
<articles xmlns="http://pkp.sfu.ca" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://pkp.sfu.ca native.xsd">

</articles>
```

XML

- **xmlTree()**: to build an internal XML trees

Challenge 5: Data frame to XML

Command:

```
aporia_xml <- function(data, file=NULL) {

    # codes continued from previous slides

    invisible(
        lapply(1:nrow(data), # nrow(data): number of rows
            function(i) {
                aporia$addNode("article", close=FALSE)
                    for(j in names(data)) { # names(data): columns name like title, abstract
                        aporia$addNode(j, data[i,j])
                    }
                aporia$closeNode()
            }
        )
    )
}
```

	title	abstract	authors	keywords
1	...	...	...	...
2	...	...	...	...

	title	abstract
1	data[1,"title"]	data[1,"abstract"]
2	data[2,"title"]	data[2,"abstract"]

Challenge 5: Data frame to XML

Outcome:

```
<?xml version="1.0" encoding="UTF-8"?>
<articles xmlns="http://pkp.sfu.ca" xmlns:xsi="http://www.w3.org/2001/
    <article>
        <title>Title1</title>
        <abstract>Abstract1</abstract>
        <keywords>Keyword1, Keyword2, Keyword3</keywords>
        <authors>Author1, Author2</authors>
        <volume>10</volumne>
        <number>1</number>
        <year>2018</year>
        <date_published>2018-01-01</date_published>
    </article>
    <article>
        <title>Title2</title>
        <abstract>Abstract2</abstract>
        <keywords>Keyword1, Keyword2, Keyword3</keywords>
        <authors>Author1 & Author2</authors>
        <volume>10</volumne>
        <number>1</number>
        <year>2018</year>
        <date_published>2018-01-01</date_published>
    </article>
    ...
</articles>
```

```
<keywords>
    <keyword>KEYWORD1</keyword>
    <keyword>KEYWORD2</keyword>
    <keyword>KEYWORD3</keyword>
</keywords>
```

```
<authors>
    <author>
        <givenname>FIRST NAME</givenname>
        <familyname>LAST NAME</familyname>
    </author>
    <author>
        <givenname>FIRST NAME</givenname>
        <familyname>LAST NAME</familyname>
    </author>
</authors>
```

Challenge 6: Reformat for keywords and authors

Command:

```
aporia_xml <- function(data, file=NULL) {

        # codes continued from previous slides

        invisible(
                lapply(1:nrow(data), # nrow(data): number of rows
                        function(i) {
                                aporia$addNode(“article”, close=FALSE)
                                        for(j in names(data)) { # names(data): columns name like title, abstract
                                                aporia$addNode(j, data[i,j])
                                        }
                                aporia$closeNode()
                        }
                )
        )
}
```

Challenge 6: Reformat for keywords and authors

Command:

```
for(j in names(data)) {

        if (identical(j, “keywords”)) {
                separate_keyword <- strsplit(data[i,j], “, “)
                aporia$addNode(“keywords”, close=FALSE)

                lapply(1:length(separate_keyword),
                        function(t) {
                                for (m in 1:lengths(separate_keyword)) {
                                        aporia$addNode(“keyword”, separate_keyword[[t]][m])
                                }
                        }
                )
                aporia$closeNode()
        }
}
```

	keywords
1	Medicine, Nursing
2	Public health
....	

separate_keyword:
[[1]]
[1]Medicine [2]Nursing
[[2]]
[1]Public health
...

separate_keyword:
[[1]]
[1]Medicine [2]Nursing
[[2]]
[1]Public health
...

separate_keyword:
[[1]]
[1]Medicine **[2]**Nursing
[[2]]
[1]Public health
...

Challenge 6: Reformat for keywords and authors

```xml
<?xml version="1.0" encoding="UTF-8"?>
<articles xmlns="http://pkp.sfu.ca" xmlns:xsi="http://www.w3.org/200
    <article>
        <title>Title1</title>
        <abstract>Abstract1</abstract>
        <keywords>
            <keyword>Medicine</keyword>
            <keyword>Nursing</keyword>
        </keywords>
        <authors>Author1, Author2</authors>
        <volume>10</volumne>
        <number>1</number>
        <year>2018</year>
        <date_published>2018-01-01</date_published>
    </article>
    <article>
        <title>Title2</title>
        <abstract>Abstract2</abstract>
        <keywords>
            <keyword>Public health</keyword>
        </keywords>
        <authors>Author1 & Author2</authors>
        <volume>10</volumne>
        <number>1</number>
        <year>2018</year>
        <date_published>2018-01-01</date_published>
    </article>
    ...
</articles>
```

Challenge 6: Reformat for keywords and authors

Command:

```
for(j in names(data)) {

        else if (identical(j, “authors”)) {
                separate_authors <- str_replace(data[i,j], “ & “, “, “)
                separate_authors <- strsplit(data[i,j], “, “)
                aporia$addNode(“authors”, close=FALSE)
                aporia$closeNode()

        }
}
```

stringr

- **str_replace()**: replace matched patterns in a string

	authors
1	Author1, Author2 & Author3
2	Author4
....	

separate_authors:

[1] Author1, Author2, Author3
[2] Author4
...

separate_authors:
[[1]]
[1] Author1 [2] Author2 [3]Author3
[[2]]
[1] Author4
...

Challenge 6: Reformat for keywords and authors

Command:

```
for(j in names(data)) {

        else if (identical(j, “authors”)) {
                # codes continued from previous slides

                lapply(1:length(separate_authors),
                        function(o) {
                                for (s in 1:lengths(separate_authors)) {
                                        name <- strsplit(separate_authors[[o][s], “ “)
                                        aporia$addNode(“author”, attrs=c(include_in_browse=”true”,
                                user_group_ref=”Author”), close=FALSE)
                                                aporia$addNode(“givenname”, name[[1]][1])
                                                aporia$addNode(“familyname”, name[[1]][lengths(name)])
                                        aporia$closeNode()
                                }
                        }
                )
        }
}
```

separate_authors:
[[1]]
[1] DAVID K. JUNG

name:
[[1]]
[1] DAVID [2] K. [3] JUNG

DAVID

JUNG

Challenge 6: Reformat for keywords and authors

```xml
<?xml version="1.0" encoding="UTF-8"?>
<articles xmlns="http://pkp.sfu.ca" xmlns:xsi="http://www.w3.org/2001/XMLS
    <article>
        <title>Title1</title>
        <abstract>Abstract1</abstract>
        <keywords>
            <keyword>Medicine</keyword>
            <keyword>Nursing</keyword>
        </keywords>
        <authors>
            <author include_in_browse="true" user_group_ref="Author">
                <givenname>Katie</givenname>
                <familyname>Lorie</familyname>
            </author>
            <author include_in_browse="true" user_group_ref="Author">
                <givenname>Tina</givenname>
                <familyname>Levesque</familyname>
            </author>
        </authors>
        <volume>10</volumne>
        <number>1</number>
        <year>2018</year>
        <date_published>2018-01-01</date_published>
    </article>
    <article>
        <title>Title2</title>
        <abstract>Abstract2</abstract>
        <keywords>
            <keyword>Public health</keyword>
        </keywords>
        <authors>
            <author include_in_browse="true" user_group_ref="Author">
                <givenname>DAVID</givenname>
                <familyname>JUNG</familyname>
            </author>
        </authors>
        <volume>10</volumne>
        <number>1</number>
        <year>2018</year>
        <date_published>2018-01-01</date_published>
    </article>
    ...
</articles>
```

Challenge 6: Reformat for keywords and authors

```
aporia_xml <- function(data, file=NULL) {
    # codes continued from previous slides
    invisible(
        lapply(1:nrow(data),
            function(i) {
                aporia$addNode("article", close=FALSE)
                for(j in names(data)) {
                    if (identical(j, "keywords")) {
                        # codes here
                    } else if (identical(j, "authors")) {
                        # codes here
                    } else {
                        aporia$addNode(j, data[i,j])
                    }
                }
                aporia$closeNode()
            }
        )
    )
    invisible(cat(saveXML(aporia_xml, indent=TRUE, file=file, encoding="UTF-8")))
}
```

R for other projects

- NCBI LinkOut using rentrez
- Digital Humanities: Text mining, text analysis, GIS and data visualization
- Data wrangling
- Data analysis
- Machine learning

Questions?

Thank you!

Yoo Young Lee
yooyoung.lee@uottawa.ca

Slide: http://hdl.handle.net/10393/38242