Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try https://etherpad.wikimedia.org).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

---

Links:

This Etherpad: https://pad.carpentries.org/2022-02-22-uw-online
Workshop website: https://lyndamk.github.io/2022-02-22-uw-online/
Library Carpentry Homepage: https://librarycarpentry.org/
Code of Conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

Lessons:

OpenRefine: https://librarycarpentry.org/lc-open-refine/01-introduction/index.html

Introductions:

[Name], [Position/Affiliation], [One thing you want to learn OR a project/task you're hoping to be able to accomplish after this workshop]

Abigail Sparling, ajsparli@ualberta.ca - Instructor, Serials Metadata Librarian, University of Alberta
Michael Lenard - Co-instructor, Thomer Lab, University of Michigan School of Information
Margarita Corral - Instructor, Data Analysis Specialist, Brandeis University

Lesley Lowery - Tech. Services Program Manager, Orbis Cascade Alliance. I'm hoping to use a combination of tools to identify all of the unreconciled OCLC merges present in our consortial catalog. I'd also like to learn more about creating collection statistics.

Daina Dickman, Helper, Assitant Director NNLM Region 5 at University of Washington. I'm hoping to learn more about R, since it is pretty new to me. Also I've been thinking about if OpenRefine can help us with some of our membership exported spreadsheets.

Amy Studer. Health Sciences Librarian, Blaisdell Medical Library, University of California, Davis. I am hoping to play around a bit with R. And, I am looking forward to hearing about work-related applications of R.

Karen Kunz, Technical Services and Systems Librarian, Oregon Institute of Technology. I really want to understand regular expressions enough to be able to create one myself from scratch.

Brian Huffman, Electronic Services Librarian, University of Hawaii School of Law Library, Get a basic handle on this core technology so I can maybe someday put it to some practical use for the library (ie learn some Ruby on Rails and set up a Blacklight instance maybe).

Jonathan Young, Science Librarian, University of Hawaii. Data skills for analyzing collections and research.

Carrie Grinstead, Providence Health, just interested in learning R!

Candise Branum, Health Sciences Librarian at Gonzaga University's Foley Library. I'm on a team tasked with providing RDM support for our faculty, and am looking to increase my knowledge around R and OpenRefine.

Ellen-Rae Cachola, University of Hawaii School of Law Library. Interested in learning data visualization skills.

Patrick Newell. Librarian at Chico State. Looking forward to learning these tools.

Lynly Beard, Health Sciences Librarian at UW - I am looking forward to learning new tools!

Lynette Teruya, Technical Svc Lib., Chaminade University of Honolulu, Looking forward to learning something new in technology

Isaac Pomper - Communication and Information Sciences PhD Student and MLIS student at UH Manoa. I have not used R or OpenRefine at all, so I'm just looking forward to learning the basics.

Caitlin Bagley, Humanities Librarian, Gonzaga University. I'd like to better understand how to clean up some data we have in our catalogue, as well as use R for the digital humanities.

David Kreiss-Tomkins, library assistant at the Sitka Public Library in Sitka, Alaska. I'm excited to learn new things!

Leah DeSantis, Active Edge and Research Services Librarian at University of Washington Health Science Library. I'm excited to know more about these data tools.

Lihong Zhu, Head of Technical Services, Washington State University.

Brian Schumacher - Systems Librarian, University of San Francisco. Have attended a Carpentries workshop before and am looking forward to learning more.

Kari Fischer - Hello! I am the Administrative and Marketing Coordinator for the Six Mile Regonal

Library District in Granite City, IL, a suburb of St Louis. I'm looking forward to learning new ways to use the massive amounts of data we have available to make better informed decisions

Andrea Lynch: I serve as the Scholarly Communication Librarian at City of Hope, a NCI-Comprehensive Cancer Center in Duarte, CA. I hope to revitalize the skills I acquired in a NN/LM data management course. I'm working to integrate some data curation/management goals into my JD.

Melissa Kahili-Heede, Librarian at University of Hawai‘i School of Medicine. Interested in re-learning Open Refine and R for research.

Nicole Gustavsen: I'm the STEM Librarian at Gonzaga University in Spokane, WA. I'm planning to offer some RDM services to my departments and thought I should learn some basics myself.

Introduction to Data

Jargon Busting

Lesson: https://librarycarpentry.org/lc-overview/03-jargon-busting/index.html

1. Brainstorm for three to five minutes on any terms, phrases, or ideas around code or software development in libraries that you've come across and perhaps feel you should know better. (Examples: linked data, OAI-PMH)
2. Next, we will assign breakout rooms of 4-6 people.
3. Make a list of all the problematic terms, phrases, and ideas each pair came up with. Retain duplicates.
4. Identify common words as a starting point - spend 10 minutes working together to try to explain what the terms, phrases, or ideas on your list mean. Note: use both each other and the internet as a resource.
5. Identify the terms your groups were able to explain as well as those you are still struggling with.
6. Each group then reports back on one issue resolved by their group and one issue not resolved by their group.
7. The instructor will collate these on a whiteboard and facilitate a discussion about what we will cover today and where you can go for help on those things we won't cover. Any jargon or terms that will not be covered specifically are good notes.

Defined/resolved

data harvesting - Retrieving data from where it's stored and moving it to a place where you can manipulate it.
data normalization - Making data consistent across files/records where it appears.
data mapping - Translating data's use/meaning/purpose in one system or process to the appropriate use/meaning/process in another system or process.
Git vs GitHub
Schema
Bootstrap
URI vs URL

sparkle - some kind of an endpoint (instructor note: SPARQL, a query language for RDF triple databases)
URI vs URL - URI is persistent
triple store - subject, object, and predicate data points that are tied together
bootstrap - a starting template, maybe includes dependencies?
Virtual Machine - server in the cloud
WAMP = Windows, Apache, MySQL, PHP
instance

Couldn't define/didn't resolve

object-oreinted program
big data
block chain
widget
data curation - esp. best practices
AIP
Fixity
lossy
package
schema
fuzzy logic
fuzzy algorithm
API+
SQL
UNIX
CSS
Ruby on Rails+
RDF and RDF Triples
BIBFRAME
linked data++
linked open data
machine learning
mirror
GNU
python
serialization

A few resources:

https://techterms.com/category/technical
https://nnlm.gov/data/thesaurus

A computational approach

Lesson: https://librarycarpentry.org/lc-overview/04-computational-approach/index.html
Is it worth the time (comic by xkcd): https://xkcd.com/1205/

Projects you'd be interested in automating:

compile staff publications
pulling statistics for journals from different publishers and vendors
statistical summaries

gate count for daily library user stats
streamline course reserves reading list processing
compiling faculty research interests - matching to library collections

automatically sending library fine reports to business office

bibliometric analysis
moving data between systems

Establish wikidata for university colleges and departments

Data cleanup

Downloading nonSUSHI compliant EResource statistics
Automating link checking
ID Management for faculty authors -- keeping track of the mulitple ways they give their name

maintaining bib list for climate center, set up form to collect data, but data is not formatted uniformly
collection development of health sciences
create reports to determine what is in catalog for certain depts
running vlookup on title matches not great
identify all unresolved OCLC merges in 11 million records (based on multiple instance of 035a and 035z)
wants to learn these tools for research. perhaps use tools for assessment. or for creating all scholarly pubs of faculty

Resources:
File formatting page: https://librarycarpentry.org/lc-overview/06-file-naming-formatting/index.html
Markdown syntax: https://www.markdownguide.org/basic-syntax/
Further reading: https://librarycarpentry.org/lc-overview/08-further-reading/index.html

Regular Expressions Introduction

Lesson: https://librarycarpentry.org/lc-data-intro/01-regular-expressions/index.html
Regex cheat-sheet: https://www.rexegg.com/regex-quickstart.html
Another cheat-sheet: https://cheatography.com/davechild/cheat-sheets/regular-expressions/

Matching and Extracting Strings

Lesson: https://librarycarpentry.org/lc-data-intro/02-match-extract-strings/index.html
Link to regex101: https://regex101.com/
Link to Markdown file: https://github.com/LibraryCarpentry/lc-data-intro/blob/gh-pages/data/swcCoC.md

Michael's regex takeaway: The expressiveness and specificity of regular expression patterns allows you

to more readily find what you want *without finding what you don't want* (what you exclude from your patterns is in ways just as important as what you include).
- e.g. if you're doing find+replace, you'll want to make sure you're only F+Ring what you intend to

OpenRefine

Lesson: https://librarycarpentry.org/lc-open-refine/01-introduction/index.html
Download dataset: https://github.com/LibraryCarpentry/lc-open-refine/raw/gh-pages/data/doaj-article-sample.csv (right click > save link as > save with a .csv extension)
Launch OpenRefine: Navigate to where you saved the program on your machine, click openrefine.exe. OpenRefine should open in your browser, if it doesn't enter this address in your browser:
http://127.0.0.1:3333/

More info on Clustering: https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

Concept Review

Faceting:

Which licences are used for articles in this file? CC BY; CC-BY-NC; CC BY-NC-ND
What is the most common Licence in the file? CC BY
How many articles in the file don’t have a licence assigned? 6

Splitting subjects:

What separator character is used in the Subjects cells? |
Split the subjects into their own cells. What steps do you take?

Clustering:

After splitting the Subject column, apply a cluster.
Which clustering method works best for subjects?
Are there any methods that don’t work?

Joining subjects:

Using what we’ve learned, join the Subjects back together

Working with Dates - SimpleDateFormat Syntax:
https://docs.openrefine.org/manual/grelfunctions#date-functions

Exercise - Reverse author names:
OpenRefine link: https://librarycarpentry.org/lc-open-refine/11-using-arrays-transformations/index.html

Going back to our facet on author we’re going to act on the cells that contain a comma by clicking on “include” for those that are true. Now that we’ve narrowed our dataset down to the lines with a comma in a name, we can use the match function. The match function allows you to use regular expressions, and

output the capture groups as an array, which you can then manipulate.

1. On the Authors column use the dropdown menu and select Edit cells->Transform
2. In the Expression box type value.match(/(.*),(.*)/) The /, means you are using a regular expression inside a GREL expression. The parentheses indicate you are going to match a group of characters. The .* expression will match any character(s) appearing 0, 1 or more times. So here we are matching any number of characters, a comma, and another set of any number of characters.
3. See how this creates an array with two values in each row in the Preview column

To get the author name in the natural order you can reverse the array and join it back together with a space to create the string you need:

2. In the Expression box, add to the existing expression until it reads value.match(/(.*), (.*)/).reverse().join(" ")
3. In the Preview view you should be able see this has reversed the array, and joined it back into a string
    - Click OK

Splitting Citation information into granular pieces

- Citation > Edit Column > Add column based on this column
- GREL: value.match(/(.*), Vol(.*), Iss (.*), \w+(.*) \((.*)\)/)[0]

Pull Data from other OpenRefine projects:

If you have another OR project it is possible to fetch and merge data.

For example, if we had a sheet with ISSN and Journal titles, we could create an additional column in our current project by pulling the Journal title in based on the ISSN data already in our data set.

To do this you use the GREL expression cell cross.

Select column you want to use as the basis to add new data and select add column based on this column

cell.cross("Project to fetch data from","Common column between projects").cells["Column to bring in"].value[0]

cell.cross("ISSNJournal","ISSNs").cells["Journal"].value[0]

Introduction to R

Lesson: https://librarycarpentry.org/lc-r/index.html
Installing and RStudio: https://librarycarpentry.org/lc-r/setup.html

Before We Start Slides: https://docs.google.com/presentation/d/1ouIBTL27vh-nG0uVcZcpSZNzWoAwh8e-i8QOPXjnmBM/edit#slide=id.p

Shortcut keys:

•	Windows	Mac
Assignment operator <-	Alt+-	Option+-
Comment/uncomment current line/selection	Ctrl+Shift+C	Cmd+Shift+C

Data Types Slides: https://docs.google.com/presentation/d/1tXxrYrrBKe7LsCl2dcwMAsX7F2W_u-_ntkChBsrQc8c/edit#slide=id.g1170b95a02f_0_23

```
rooms_no_na <- rooms [!is.na(rooms)]
rooms_no_na
rooms_no_na2<- rooms[complete.cases(rooms)]
rooms_no_na2

median(rooms, na.rm=TRUE)

room_above_2 <- rooms_no_na[rooms_no_na>2]
length(room_above_2)
```

DATA:
Go to the Figshare page for this curriculum and download the dataset called “books.csv”. The direct download link is: https://ndownloader.figshare.com/files/22031487. Place this downloaded file in the data/ you just created. *Alternatively*, you can do this directly from R by copying and pasting this in your terminal:

```
download.file("https://ndownloader.figshare.com/files/22031487","data/books.csv", mode = "wb")
```

Resources to learn R

Parts of this episode have been inspired by the following:

- “Before We Start” R for Social Scientists Carpentry Lesson. CC BY 4.0.
- Roger Peng’s Computing for Data Analysis videos
- Lisa Federer’s Introduction to R for Non-Programmers
- Brad Boehmke’s Intro to R Bootcamp

Cheat Sheets

- **Base R Cheat Sheet by Mhairi McNeill**
- **Data Transformation with dplyr Cheat Sheet by RStudio**
- **Data Wrangling with dplyr and tidyr Cheat Sheet by RStudio**
- **Complete list of RStudio cheatsheets**

Data cleaning & transformation with dplyr

Lesson: https://librarycarpentry.org/lc-r/03-data-cleaning-and-transformation/index.html
Free book for using Tidyverse for data analysis: https://r4ds.had.co.nz/
If you aren't already set up from the previous section:

1. File > New Project. Create a new project with an appropriate name (library_carpentry, etc.).
2. File > new file > R script. Save the script as script.R, library_script.R, etc.
3. Paste then run the following code into your script:

```
library(fs)
dir_create("data")
dir_create("data_output")
dir_create("fig_output")
download.file("https://ndownloader.figshare.com/files/22031487",
        "data/books.csv", mode = "wb")
library(tidyverse)
books <- read_csv("./data/books.csv")
```

4. Check to see if the books object is showing in your environment window