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 etherpad.wikimedia.org).

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

Chuck McAndrew - IT Librarian, Lebanon Public Libraries, NH, Chuck.mcandrew@leblibrary.com, walked across Spain
David Yakobovitch - Principal Data Scientist, Galvanize david@yakobovitch.com, Visited South Korea this year

Name, Current Title/Function/Role, What Institution/Organization, Email/twitter, What you want to take away from the workshop, Fun fact that one of your Best friends know that is G-Rated

Diana Reid, Electronic Resources/Discovery Librarian, University of Louisville,diana.reid@louisville.edu, Wore mittens as socks
Rebecca Pattillo, Metadata Librarian, University of Louisville, rebecca.pattillo@louisville.edu, better understanding of SQL, regular expressions, and data manipulation to expedit my work, Fun Fact: before being a librarian, I was a flight attendant
Terri Holtze, Head of Web Services, terri.holtze@louisville.edu, improve programming skills, can dance tango
Trish Blair, Collections Manager, University of Louisville, trish.blair@louisville.edu, to learn the processes of data management, I like to put hats on my cats
Heather Fox/Director, Oral History Center, Manuscript Archivist/University of Louisville/heather.fox@louisville.edui want to be able to manipulate data more efficiently and to clean up metadata in the OHC database with open refine/I just started playing Dungeons and Dragons with my daughter [who is DMing for the first time] and her nerd friends. I am a golaith warlock with a skeletal hand.
Brandon Harwood, Project Coordinator, UofL, brandon.harwood@louisville.edu, which programs can do what, I always carry a hand fan with me.
Christopher Poche - Ekstrom Library - Library Specialist, Current Periodicals & Microforms - this is new to me, so just something new to help me see new possibilities in my work - was an impromptu antiques salesperson in Florence, Italy
Rachel Howard, Digital Initiatives Librarian, University of Louisville, rachel.howard@louisville.edu, tech skills to assist with batch data migrations, I can walk on stilts
Jen Montavon-Green, Electronic Resources Manager, University of Kentucky, jen.montavon-green@uky.edu, hoping to learn some better ways to manage the data I work with, I lost $22K on Jeopardy thanks to leaving one letter out of my answer (I couldn't remember if it was Larry McMurtry or McMurty. I guessed wrong.)
Mary Beth Williams, Reference Associate, University of Louisville Archives & Special Collections, mbwill07@louisville.edu, all the info! this is all new to me, knitting wizard

Amber Willenborg, Online Learning & Digital Media Librarian, University of Louisville, arball02@louisville.edu, general understanding of data tools, hates ketchup
Leah Cover, Instruction Librarian, Spalding University, lcover@spalding.edu, To gain a basic understanding of data tools and how they can serve my/our work as librarians. There is a mouse named Cookie living in my house.
Angel Clemons Electronic Resources Librarian angel.clemons@louisville.edu new skills to make converting data I've collected into a more accessible database Fun Fact: I love the outdoors
Anita Hall, Assessment & Analytics Librarian, University of Louisville, anita.hall@louisville.edu, I'm hoping for a refresher on some skills I haven't used in a while and to get some hands-on practice with new tools, I am also a registered yoga teacher
Laura Edwards, Team Leader-Discovery and Access, Eastern Kentucky University, laura.edwards@eku.edu, better understanding of data management/manipulation tools; better able to communicate with systems/tech colleagues. Fun fact: I enjoy reading and learning all the things and hiking
Casey Jett, Library Specialist, JCTC, casey.jett@kctcs.edu, learn about how these technologies work in libraries, I visited Supai, Arizona in October which is where Havasupai Falls is located
Ken Solomon, Library Coordinator, JCTC, ken.solomon@kctcs.edu, better understanding of data management. Fun fact: I have two cats, Dhango and Josie.
Courtney Baron, Director of the Art Library, University of Louisville, courtney.baron@louisville.edu, looking for a refresher on concepts learned in graduate school (SQL, database design, etc.), I just got engaged at Disney World!
Jason Zahrndt, Digital Media Consultant, University of Louisville, jason.zahrndt@louisville.edu, better understanding of data managament, favorite book is The Last Unicorn
Danielle Spalenka, Digital Collections Curator, The Filson Historical Society, danielle@filsonhistorical.org, learning new tech skills that may be applied to my job (metadata management; digital preservation) Fun fact: I was in a bagpipe band in college
Michael Losavio, Department of Criminal Justice , quatro filhos, hope to get a handle on these technologies
Ariel Shirley, Part-Time Circulation Assistant, Spalding University, ashirley01@spalding.edu, a better understanding of how to utilize data in the workplace, I love reading comic books
Sarah Frankel, Open Access & Repository Coordinator, University of Louisville, s.frankel@louisville.edu, learn new skills and software, Fun Fact: I love metal music!
John Merritt, Director of Library Technology, SBTS, jamerritt@sbts.edu, new (data wrangling) skills, I'm currently reading LOTR with my son.
Rebecca Jude, E-Resources Specialist, University of Kentucky, rebecca.jude@uky.edu, better understanding of data management tools that can be applied to current position, Fun Fact: I went to Bora Bora in March 2019
Kyna Herzinger, Archivist for Records Management, University of Louisville, kyna.herzinger@louisville.edu, new info to help me do my job better, love watching William Powell movies
Amy Purcell, Associate Curator, Archives & Special Collections, Univerity of Louisville, apurcell@louisville.edu, assisting users local and remote to access our collections, Son is at CERN
Mimi O'Malley, Spalding University Sr. Mgr., Library and Online Services,momalley@spalding.edu, overview of data ; write 3 books
Fannie M. Cox, Outreach & Reference Librarian, UofL, fmcox@louisville.edu, just interested, fun fact- I like the Matrix.
Lidiya Grote, Social Sciences Librarian, lidiya.grote@louisville.edu, went to Ukraine this summer
Sarah Drerup, STEM Librarian, sarah.drerup@louisville.edu, qualified expert on the .50cal this year
Riley Sumner, Hospital Librarian, University of Louisville, riley.sumner@louisville.edu, just interested in

learning new skills, huge Harry Potter fan

Jargon Busting
regex/regular expressions
Versioning/version control
normalizing
SQL
API
AI
Big Data

Regex Notes
Use https://regex101.com/ to test your regex
Cheat Sheet at https://librarycarpentry.org/lc-data-intro/reference.html
https://github.com/LibraryCarpentry/lc-data-intro/tree/gh-pages/data/swcCoC.md
organise
orgenize
organize
organife
organized
Organise
disorganized
Organize
organize a strike
you should organize a strike

RegEx use cases?
- Open Records requests, searching for keywords
- Take data from word to excel and reformulate, helps with placing information correctly

"\s" White space includes spaces, tabs, new line characters - Is there a way to search for double tabs? Yes: \s{2}
"\S" opposite, matches non-whitespace (capital letter frequently indicates "not")
"\b" "word boundary" will find the start/end of a word. looks for space, period, dash, etc. - but NOT underscore
"." matches a single character. add + to match one or more ".+" add ? to match 0 or 1 ".?"
"|" is an or (pipe symbol)

"in the wild" how would you use RegEx?
- depends on environment - could use word or excel, allows regex in find/replace function, python or pearl script,
Since you can search for keywords, is there a way to use this in library instruction when showing students how to use databases?
- Python, SQL, pearl scripts
- Any time you want to automate something
- Search and replace functions in lots of different software, languages.

Fr[ea]nc[eh]
Fr[ea]nc[eh]$

\b\S+@\S+\b
\w[@]\w*[.]\w{3}

Question: How to match all email addresses except those ending in .com
Answer: [\w\.-]+@[\w\.-]+\.(?!com)\w+\b
(?!com) is called a negative lookahead and means the the next expression will be discarded if it matches com"

Often more than one way to write an expression that will work, in theory you want it to be as concise and efficient as possible

https://librarycarpentry.org/lc-data-intro/01-regular-expressions/index.html

More Challenging Regex Problems:
https://regexcrossword.com/

SQL
.sql files generate scripts
Can create/delete/manipulate databases with SQL as well as query
Returns a table of results, rather than list of matches
Databases contain Tables which contain data/metadata in rows and columns (can be huge) - Relational when they have common elements. Several smaller tables are faster, easier to work with than one giant table. Also good for data normalization / ensuring that data is consistent.

SQLite is most commonly used version

Visual studio code (text editor that is easier to read)
https://code.visualstudio.com/
MySQL Dump - backs up your DB and creates SQL structure for import

query is a request for data like regex was a request to find patterns

SQL Commands: SELECT column FROM table (command doesn't have to be capitalized, but good practice for readability, but software may be case-sensitive for column/table names)
SELECT title
FROM articles;

SELECT what
FROM where
(Additional Logical Statements)

Style notes: (there is some variability between versions)
Commenting out a line: "--" double hyphen for single line or for multiline /* to start and */ to close
Indentations and line breaks are ignored, can use to make things more readable
Separate multiples with commas
* selects all
% is a wildcard (0 or more characters)

Placing logical statements in () can help keep them separate and not interfering with each other.
Logical ordering matters! X OR Y AND Z - how does it parse? (X OR Y) AND Z versus X OR (Y AND Z)

DISTINCT will eliminate duplicates - works only on the immediately following value, there are more complex ways to use it on multiple values? Unclear. This is a "Keyword" which is used to run logical statements.
ORDER BY to sort
WHERE to filter (or HAVING to filter on an aggregate function in place of WHERE)
LIKE, IN, BETWEEN AND, IS NULL
REGEXP 'expression'
AS to rename columns
IN ('x', 'y', 'z') can replace OR

Aggregate Functions (like pivot tables):
AVG, MAX, MIN, COUNT, SUM (not a math language so there is a limited number, but you can build)
by default, aggregates for all results, but can set groups with GROUP BY
Rename with AS so that you can call that calculation more easily

Custom Calculations, can build/wrap

JOIN to combine two tables ON a common column. Additional JOINs happen one at a time

What do you want to ask, how will you order commands, what logic will you use?

By default, SQL just views data. Can INSERT INTO table or create a new table (easier). For numbers that could change over time, would want to save the query instead and rerun as desired.

Real world example of database schema http://schema.koha-community.org/19_05/

```
SELECT
    a.ISSNs,
    j.Journal_Title,
    round(avg(a.Author_Count),2) as avg_author_cnt
FROM articles as a
JOIN journals as j
ON a.ISSNs = j.ISSNs
GROUP BY a.ISSNs
HAVING avg_author_cnt > 6.52
ORDER BY avg_author_cnt DESC
LIMIT 10;
```

Joining multiple tables:

```
SELECT a.Title, a.First_Author, j.Journal_Title, p.Publisher
FROM articles as a
JOIN journals as j
ON a.ISSNs = j.ISSNs
JOIN publishers as p
ON p.id = j.PublisherId;
```

INNER and OUTER joins (like a venn diagram) - INNER is default. LEFT OUTER and RIGHT OUTER (can help pull in data rows where info is missing)

VIEWS
CREATE VIEW view_name AS (rest of query), can then call the query as desired
SELECT *
FROM view_name;
DROP VIEW view_name; to delete

https://librarycarpentry.org/lc-sql/

http://www.pyregex.com/

Open Refine Material:
https://librarycarpentry.org/lc-open-refine/
https://librarycarpentry.org/lc-open-refine/setup.html

Split multi-valued cells: Choose arrow next to column header > edit cells > split multivalued cells
- splits into multiple rows, switching to records mode will show them grouped together
- can re-join with Join multi-valued cells (choose separator - pipe is a nice safe uption because not used in other written contexts)

Facet > Text Facet will treat as strings. Basically a query of your data.
- all based on True/False
- can then edit the string to apply to all that match
- other types of facets, including facet by errors/null/blank/etc. Note: null is usually numeric data, empty string text.
- can do multiple facets and they will interact with each other

Important to remember: Undo/Redo history - if you make a new split, you lose the history of the previous one
- Can use "extract" to copy/paste JSON of steps (like a macro)

GREL: General Refine Expression Language (similar to RegEx) powers OpenRefine
- same syntax as Python
- you can use other languages if desired

Cluster and Edit: find similar strings and group together (method to find and normalize errors).
- different options for algorithms/sub-algorithm available
- Can merge and standardize..."Browse this cluster" to check

Data Transformations: functions. Can write custom, use common transforms in OpenRefine, use online resources to find what you're looking for.
- some functions require specific data types, ex: can convert to date and then have all date functions available

This is just a view, not changing the underlying data file. If you want to use the new cleaned-up data, you

can export it out.
- Can also retrieve from a URL to connect into an API
- Could export to a SQL table, etc.

https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language

https://librarycarpentry.org/lc-open-refine/10-data-transformation/index.html

What is "bash"? - it's a command shell, standard for linux/unix type shells. uses certain commands that aren't covered in this course, but you can look them up

"git" is the software, "github" is a site that hosts git repositories (now owned by Microsoft - offers free/paid, public/private)
- you could use git without ever using github, or use a different hosting site, etc.
- lots of popular open source software, including library software, is on github (i.e. Omeka)
- there are also non-code projects on github (ex: library policies document in markdown)
- can set up collaborators/organizations within github. non-collaborators can make a "pull request"

git add to place into staging area
git commit -m 'message text' to save to repository
- should answer the question "If I commit this, I will...."

when you are finished working, check git status. want to see "nothing to commit, working tree clean"

most people actually use command line for this - there is a graphical version but may find it more confusing
- some text editors have ability to commit on save built in

Setting up a repository in github - if you add a README.md file, github will display the context on the repository page (including license!)

git remote -v will show you the url of your repository

steps:
- edit file with text editor (save/close)
- git add (filename)
- git commit -m 'message text'
- git push -u origin master (after first time can just use git push, it will keep using same branch until you change it)

when collaborating with someone else, copy repository address and git clone (url)
- look for directory name and change directory cd (directoryname)
- if not sure where you are, pwd (print working directory)
cd .. goes back up a level

Creating branches...i.e. production and testing
git merge when you are ready to put them back together

From an analysis perspective, python is good language to learn/know

Interested in more openrefine training? Check out KCA Spring Meeting May 14-15
www.kyarchivists.org

Extra Resources:

https://github.com/davidyakobovitch/data_science_resources
https://pypi.org/

https://chuckmcandrew.com/2019-12-12-louisville/

https://librarycarpentry.org/lc-git/setup.html

https://choosealicense.com/

Git command tips: http://try.github.io/

https://github.github.com/training-kit/downloads/github-git-cheat-sheet.pdf
https://librarycarpentry.org/lc-git/reference