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

All content is publicly available under the Creative Commons Attribution License: https://creativecommons.org/licenses/by/4.0/

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

All content is publicly available under the Creative Commons Attribution License:https://creativecommons.org/licenses/by/4.0/

Our workshop! Schedule, syllabus, etc.
https://smithsonianworkshops.github.io/2020-09-15-Smithsonian/

Pre-workshop survey! Please do this now, if you haven't yet.
https://carpentries.typeform.com/to/wi32rS?slug=2020-09-15-Smithsonian

* **Day 1:** Introduction to Working with Data

Intro slides: https://docs.google.com/presentation/d/1ijK6WjPbnvupF7OcjKIdCFXuYA_CMXW7-rSR2kpuHTY/edit#slide=id.g989ba3a7e5_0_132

Introduction to Working with Data lesson: https://librarycarpentry.org/lc-data-intro/

Instructor: Matthew Kweskin, NMNH

* Attendance Day 1
Dana Feil, SLA, Libraries, Eresources & Serials :)
Name, Unit, and your favorite emoji (😀) *https://emojipedia.org/*
Stefaan Hurts, SLA-DPI
Alliw Alvis ♥😂
Daniela Jiménez, NMAH, 😒
Daniel Euphrat, SLA
Carrie Smith, STRI, \w/
Heidy Berthoud, SLA
MItch Toda, SLA, Archives and Information Management :)

Stephanie Kurasz NMAH
Ricc Ferrante, SLA Digital Services, Archives ;-)
Sue Zwicker SLA Research Services- SERC/NH [Caprentries helper / lurker :-)]
Suzanne Pilsk, Smithsonian Libraries and Archives, DPI Metadata Department
Manuel Samudio, STRI,
Eden Orelove, OCIO
Sue Graves, Smithsonian Libraries and Archives, DPI Metadata Department😎😎
Alex Edezhath, Discovery Services, SLA
Richard Naples, SLA ♂
Erik Bergstrom, SLA
Jackie Chapman, SLA
Bonnie Felts, SLA
Carrie Craig, NMNH,
Alison Oswald, Archives Center, NMAH
David Holbert, SLA
Mike Trizna, OCIO,
Matthew Kweskin, NMNH,
Amanda Devine, NMNH,
Bess Missell, SLA, 😊
Laurie Stepp, NPG

Goals:
Know what it is
Demistify
Incorporate it in your work

*Resources:

- Regex LC website: https://librarycarpentry.org/lc-data-intro/
- Cheat sheet good for what is covered in this lesson: http://web.mit.edu/hackl/www/lab/turkshop/slides/regex-cheatsheet.pdf
- Online problems and tutorial: https://regexone.com/
- Online regex system we'll be using: https://regex101.com/
- Online regex quiz (requires registration): http://regex101.com/quiz

*Source material:

- https://acrl.ala.org/techconnect/post/fear-no-longer-regular-expressions/
- Understanding Regular Expressions, Damian Conway: https://www.oreilly.com/library/view/understanding-regular-expressions/9781491996300/ (behind pay wall)
- https://www.regular-expressions.info/
- https://en.wikipedia.org/wiki/Regular_expression

*How Do You Use Regular Expressions?
*Mac:
BBedit
Sublime Text
Visual Studio Code

*Windows:
Notepad++
Sublime Text
Visual Studio Code

*Flavors of Regular Expressions
PERL Compatible Regular Expressions (PCRE) - what is most commonly used, though there are other varieties out there, so check the documentation in the program you are using to see the specifics for that flavor of regex.

https://regex101.com

*common regex

ranges of characters:
brackets - [] you can put in specific ascii characters in there, case sensitive
\w matches any word character (the same as [a-zA-Z0-9_] (this doesn't include things like diacritics)
\W (capital W) does the opposite of \w - it matches any non-word characters (including diacritics)
\d matches any digit
\D matches any non-digit
\s matches white space characters (spaces, tabs, returns)
\S matches non-whitespace characters
adding a carat, ^ in your bracket will exclude those from the search and find everything else

quantifiers:
in regex101.com, there is a quantifiers section in the quick reference that explains many of these: adding a plus sign + after your brackets matches between one and unilimited times--it's called greedy because it keeps going. adding ? after your brackets matches between zero and unlimited times.

question:
--great question to figure out! How do you search for characters with diacritics?

escaping:
You can "escape" special characters by adding a \ before the character: so to find a question mark, search \? instead of just ? since ? is a reserved character

anchors
anchors can match the beginning and ending of a line. Adding the carat at the start means that it will only find lines that begin with your search, while the dollar sign finds only ending of lines that match

others
\b specifies that the match has to be after or before a word break or boundary--a space or a line return.
\s finds a space

options
m = multiline. Each line in the document has its own beginning and its own end. If that option is not included, all the text will be treated as one line
g = global. All matches in the document will be found. If this is off, only the first match will be found.
i = case insensitive
Text editors, other software, etc. usually has a checkbox for these options.

Exercises

1 What would match the strings French and France that appear at the beginning of a line?
^Fr[ea]nc[eh]
^[F]r[ae]nc[eh]
^Fr[ea]nc[he]
^Fr.nc.

2 How would you find the whole word headrest and or head rest but not head rest (that is, with two

spaces between head and rest?
\bhead ?rest
head.?rest
head\s{0,1}rest
head[\s]?rest
\bhead ?rest\b
head\s?rest
\bhead\s?rest

3 How would you find a string that ends with four letters preceded by at least one zero?
\b0+[a-zA-Z]{4}
^[0]{1,}[A-Za-z]{4}$
0+[A-Za-z]{4}
0[a-zA-Z]{4}\b

4 How would you match the date format dd-MM-yyyy?
[0-9]{2}-[0-9]{2}-[0-9]{4}
\d{2}-\d{2}-\d{4}

5 How would you match publication formats such as British Library : London, 2015 and Manchester University Press: Manchester, 1999?
([A-Z][a-z]+\s?)+:[a-zA-Z\s]+,\s[0-9]{4}

[A-Z][a-z]+\s? inside the parentheses matches a word with one capital letter, at least one lower case letter, and possibly a whitespace character. **()+** takes this whole expession and says that there should be at least one, but maybe more, of these words.

SQLite does not support full PCRE regex, but it does have some wildcard characters you can use with SELECT statements: https://www.sqlitetutorial.net/sqlite-like/

Matching & Extracting Strings: https://librarycarpentry.org/lc-data-intro/02-match-extract-strings/index.html

Example text: https://github.com/LibraryCarpentry/lc-data-intro/blob/gh-pages/data/swcCoC.md

Pre-filled RegEx101 link: https://regex101.com/r/TlRsxY/1

If you want to really nerd out on email regular expressions, check out: https://www.regular-expressions.info/email.html

Matthew Kweskin kweskinm@si.edu

Daily Feedback Survey: https://docs.google.com/forms/d/e/1FAIpQLSckvNc-MCKaW11B55YYAYhPS-OYRawA8RIJugDeG4VoyUX6mw/viewform
----

*Day 2: Tidy data for librarians

Tidy data for librarians setup: https://librarycarpentry.org/lc-spreadsheets/setup.html

Tidy data for librarians lesson: https://librarycarpentry.org/lc-spreadsheets/

*Attendance Day 2

Please add your name and your favorite thing you learned about the Introduction to Working with Data (yesterday's lesson or a question you have:
Stephanie Kurasz, working in regex101
Carrie Craig, seeing the debugging feature at regex101.com
Alison Oswald--learning about regex 101 and its power. It's going to take some time to learn all of the commands
Alex Edezhath -- first time learning about regex (and asking my colleagues how they use it)
Amanda Devine - I didn't know how to use \b before!
Daniel Euphrat - seeing multiple ways to accomplish the same search
MItch Toda - Seeing how regex can do complex searches
Eden Orelove - learning some practical applications (uses) of regular expressions
Sue Graves, working in Regex101
Carrie Smith, learning about regex
Suzanne Pilsk, regex101
Allie Alvis, so much RegEx!
Manuel Samudio, regex is great!
Richard Naples, regex101.com was the bomb dot com
Jackie Chapman, regex101 as a place to practice/learn/check
Daniela Jiménez, Regex101 and learning about the variety in developing expressions
Stefaan Hurts - discovering there was something like regex!
Heidy Berthoud, regex101 website
Regex to find characters at the end of a line you may may want to remove: [\s;,.]+$
This finds any punctuations (; , .) and "whitespace" characters (with \s: spaces, tabs etc.)
Ricc Ferrante - what Richard Naples said!
David Holbert - RegEx
Bonnie Felts - RegEx (I had no idea what this even was, so it was a cool resource)
Bess Missell - regex101 - very helpful!
Laurie Stepp - Regex 101 is invaluable, I use to extract data from .xml sidecar files for AV

BREAKOUT ROOMS:

- **How many people have used spreadsheets in their work?**

- Most people use spreadsheets - some daily, some less often

- **What kind of operations do you do in spreadsheets?**

- Pull data out of systems in csv, work/clean, and load back
- Macros
- Data cleaning
- Working with library data
- Tracking work, reporting work, sharing work
- Tracking collections data
- Extracting data
- Performing statistical calculations
- Tracking usage data
- Tracking budgets
- Working with data from the DAMS, ArchivesSpace
- Concatenating data

- **Which ones do you think spreadsheets are good for?**

- Conditional formatting
- Cleaning and changing data
- Looking at tabular data, seeing ranges of values
- Being able to see all data at once
- Functions

- **Spreadsheets can be very useful, but they can also be frustrating and even sometimes give us incorrect results. What are some things that you’ve accidentally done in a spreadsheet, or have been frustrated that you can’t do easily?**

- combining data - columns from other places to mush together and deduplicate
- Excel auto-formatting of barcodes, call numbers, dates
- Excel shortens really long numbers into #######
- Copying and pasting data (especially with formulas) can have unexpected results
- Ordering can be tricky
- When importing data into Excel, Excel automatically creates a table instead of just importing the data into plain cells
- Permissions can be tricky
- You can get lost when trying to work in a very large spreadsheet. They can get unwieldy.
- Bulk editing can be tricky
- It can be hard to see extra spaces, hidden unicode characters, etc.
- Excel can sometimes be too helpful and suggest or take actions that you don't want
- Receiving data that needs to be formated to be read by excel

https://librarycarpentry.org/lc-spreadsheets/data/training_attendance.xlsx

Tidy Data diagram (from R for Data Science):
https://r4ds.had.co.nz/tidy-data.html#fig:tidy-structure

The cardinal rules of using spreadsheet programs for data:

1. Put all your **variables in columns** - the thing you’re measuring, like ‘length’ or ‘attendance’.
2. Put each **observation in its own row**.
3. **Don’t combine multiple pieces of information in one cell**. Sometimes it just seems like one thing, but think if that’s the only way you’ll want to be able to use or sort that data.
4. **Leave the raw data raw** - don’t mess with it!
5. Export the cleaned data to a **text based format** like CSV. This ensures that anyone can use the data, and is the format required by most data repositories.

Ways that this data is not "tidy":
column D has more than one type of data - What the heck "Other"
c & h - hours
Column C doesn't display decimal points consistently
No definition of initials in columns E and J
Rows are in two tables are not the right
Can column D be split inito three separate?
date value formatting forces you to look at the tab label to figure out the year

Tidy Data Publication: https://www.jstatsoft.org/article/view/v059i10

BREAKOUT ROOMS: LIST FORMATTING ISSUES
(You have until 10:35)

Group 1

- widen columns, different formatting (general, custom, date)
- alignment issues
- knowing the different excel paste options
- combine tabs 2016 +2017? or keep separate? depends on project
- use of acronyms
- dates are very frustrating!
- search and replace not always working

Group 2

- 2016 - Hours - consistent format, use of shading to indicate a cancelled event - should be a separate column
- 2017 - Inconsistent date format /use of ?, multiple variables in column D - should be broken out into separate columns, special character use in column E, empty cells in columns C and D, use of shading to indicate a cancelled event - should be a separate column, two dates in one cell
- Dates - Hours: need to be in a consistent format, Num_registered first entry is 1.5 - possible data entry error, Cancelled column - change to numeric indicator
- Combine data into one set, with a column to indicate whether RDM or Open access training, include also columns for registered and attended
- Date formatting - Change to numeric so that you can sort
- Delivered by - Use full name - Last name, first name instead of initials incase there are duplicates initials

Group 3

- -inconsistent length columns; even when length in hours is specified, this format is not followed (60 hour training??)
- -Needs consistent headings and measurements (hours? mins?)
- -Initials are not a great way of tracking, especially if only by two initials. Add column for last name (at least). Also, there's at least one training that was delivered by multiple people, so we need a way to represent that.
- -Inconsistent date formatting; on the "Dates" tab, it's impossible to tell what year the training took place in without detective work. On the 2017, there is different date formatting within the same table, some with "?"
- -The PGR/PDRA/Other column really needs to be three separate columns. Would be nice to spell things out in the headings, in case people unfamiliar with the abbreviations ever look at this sheet.
- -The tables on the 2016 and 2017 tabs need to be on distinct tabs; should not live side by side like this
- -Resolve the "cancelled" workshops, perhaps on their own tab? Every tab deals with canceled differently.
- -Some tables have blank cells, which could cause problems.
- -More thought needs to be put into categories; the type of data that is tracked changes over time (see 2016 OA training vs. 2017, table records both registered and attended at a later date).
- -Has this group decided what's important to them? Would they rather present training by year, or by training type? Is it important to highlight trainings that occurred or trainings that were canceled? What is the purpose of this data, and how is it serving them?

Group 4

- Inconsistent dates
- Include year in dates (dont rely on tabs)
- Row 2 data should be a column (RDM / OA)
- combine year tab data into one set of data
- cancelled should be a column (Y/N or T/F)
- combine attendees- PGR / PDRA / Other / Unknown
- delivered by - all depends on end use case, but especially here - two columns? Or trainers T/F?
- two dates in 2012 B5
- uncertainty in 2017 B11/12
- Registered... we have this data sometimes but not other times
- Dates tab
- - D2 - how can half a person be registered?
- - type has OA and RDM and... Other?
- - time in hours, some of these in minutes
- - no year...?
- so many problems
- tabs should probably be combined, but don't all have the same fields which would lead to blank cells in the combined version

Questions re: being tidy vs being accurate. <-- and knowing what your end goal is and how that informs formatting etc YES and also trying to anticipate future needs that you do not need yet.... but also not doing that so much such that you never get anything done? (perfect getting in the way of the good)
garbage in garbage out. (not trusting data like attendees/registered etc)
This made us all sad!
Haha. I know we didn't talk about it, but do column headers have to have a specific format?
What about unique IDs?
If you want to make your spreadsheet the most machine readable, it's good to have headers without spaces and non-alphanumeric characters. Generally you want to stick with letters, numbers, digits, and underscores. You can also have problems if your header starts with a digit. That said, a lot of software can deal with column headers that do include spaces, characters, etc.
For unique IDs, if you can control it, I would probably not have IDs that start with 0 (since Excel loves to truncate leading zeroes), and I might not have IDs that could be misinterpreted as dates.

Group 5

- EGREGIOUS
- 2016, columns c & h - should have 'hours' in heading
- column d, breakout into 3 headings
- cancelled in L7
- shading?
- RDM Training / Open Access headings -- make 'own columns -- see Dates tab (renamed 2015)
- rows don't reflect same date
- what does Other signify?

Group 6

- --column d (make three columns and distribute values accordingly
- --column i are not the same way split as column d both are attendees)
- -Add notes that identify the instructors to the corresponding initials
- -Breakdown attendees' categorization (PGR, PDRA, Other) in OA training as well
- -Date format
- -Consistent use of decimal numbers and place value (2.0 vs 2) for length of the instructions, time should be tracked the same way (hours or minutes?), unit measure should be identified in the column header, not in the unit value

- -Equivalent column headers should have the same column header label (Length, not Len)
- -"Dates" tab does not keep track of the year the data corresponds to
- -"Cancelled" events are highlighted in tabs 2016 & 2017 but in tracked through a separate column under "Dates" tab
- -If years were included in 2016 and 2017 tabs, data could be combined under one spreadsheet
- -Shouldn't have two tables on same spreadsheet
- --attendees vs. registered (cancelled classes have "attendees")

For naming columns, I love all the different cases out there like Camel Case, Snake Case, Kebab Case, etc: https://medium.com/better-programming/string-case-styles-camel-pascal-snake-and-kebab-case-981407998841
^ "Which is best? There is no best method of combining words. The main thing is to be consistent with the convention used, and, if you're in a team, to come to an agreement on the convention together." Yes! +1
Along these lines, I think different data standards and style guides will have guidelines for how variables/ column names should be formatted.

Data is side by side that could skew data sorting
No headers that are set at top to organize/sort data

https://i2.wp.com/res.cloudinary.com/syknapptic/image/upload/v1521304412/messy_tidy_qq4ba9.png

https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates
Was going to post link to this piece but glad to see it's already here. Great case study.

Smithsonian has a workshop on pivot tables that is recommended: go to moodle.si.edu on the network and search. Or open this link while you're on the network and sign in:
https://si-vmerplms.si.edu/moodle/enrol/index.php?id=1620

In Excel, the TSV format is called Tab delimited Text (*.txt). Although the file is saved as a text file, the values will be separated with tabs and can be worked with as though it were a TSV file.

Conditional Formatting tips:

- when sorting, you can sort a column by color. This can be useful if you want to move all of a particular color to the top of the list to work with those records
- use conditional formatting to identify duplicate values in a column

ISO 8601 format for dates is best! (https://www.w3.org/TR/NOTE-datetime)
This shows the year first, then month, then date, separated by dashes: 2020-09-16 is the date of our workshop. It's international format, and will be interpreted by excel and other computer programs. (America HAS to be special...)

Best practice, though, is to separate year, month, and date into their own columns. To do this in Excel when you already have a full date, is to add the formulas =YEAR(), =MONTH(), =DAY(), with the cell you want to convert put into the parentheses.

You can also get the month name instead of the number.
https://exceljet.net/formula/get-month-name-from-date
=TEXT(A2, "mmm")
=TEXT(A2, "mmmm")

From Wikipedia (https://en.wikipedia.org/wiki/Year_1900_problem):
Microsoft Excel (using the default 1900 Date System) cannot display dates before the year 1900... Excel uses a floating-point number to store dates and times. The number 1.0 represents January 1, 1900, in the 1900 Date System, or January 1, 1904, in the 1904 Date System and was the default for Macintosh prior to Excel 2016. Numbers smaller than this display as a #VALUE! error.

To import text data into Excel:
Data > From Text/CSV

Daily Feedback Survey: https://docs.google.com/forms/d/e/1FAIpQLSckvNc-MCKaW11B55YYAYhPS-OYRawA8RIJugDeG4VoyUX6mw/viewform

----

*Day 3: OpenRefine

OpenRefine setup: https://librarycarpentry.org/lc-open-refine/setup.html Any version of OpenRefine from 3.2 on will work. I will be using 3.4

OpenRefine data set: https://github.com/LibraryCarpentry/lc-open-refine/raw/gh-pages/data/doaj-article-sample.csv

OpenRefine lesson: https://librarycarpentry.org/lc-open-refine/

GREL functions for more complex data transformations:https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions

Notes:
If you open OpenRefine, and it doesn't automatically open a window in your default browser, try one of these:
http://127.0.0.1:3333/
http://localhost:3333

Don't use Edge, nor Internet Explorer. It does work in Chrome, Firefox, and Safari

*Attendance Day 3

Please add your name and your favorite thing you learned about Tidy data for librarians:

Suzanne Pilsk - Everyone at SLA should understand what Tidy data is and best practices so we can stop the madness +1
Alex Edezhath -- there ARE standards and best practices for data on a spreadsheet/csv/tsv :)
Allie Alvis-Seconding Suzanne!! It would be great to have a required workshop on tidya data...
Bess Missell - the limitations on dates! +1 (Suzanne wow, that was news!)
Eden Orelove - strategies for working with dates
Daniela Jiménez - documenting 'null data'
Daniel Euphrat - the differences between "tidy data" and the common sense understanding of what makes data neat/unmessy
Alison Oswald

Stefaan Hurts - tidy and messy are in the eye of Excel and not other beholders
Richard Naples - How to deal with dates
Manuel Samudio -
Mitch Toda - The complexities of data and formating. Also the importance of starting well at the beginning on how data should be entered.
Sue Graves - Tidy data means having an eye toward the larger picture from the begining
Jennifer Giaccai - not having variables in column names. I have two tables I work with that I'm doing that right now--oops!
Laurie Stepp - keeping clarity and working with dates
Stephanie Kurasz
David Holbert
Heidy Berthoud - Know the limitations of Excel, and take your time setting up spreadsheet.
Bonnie Felts - "Null value" vs 0 vs 9999 (or whatever else) could really mess with your data.
Carrie Smith - <Etherpad working in Edge> - all of the everything to better work with the library data from an unsupported OPAC...

Java is now bundled with OpenRefine for 3.4: https://openrefine.org/download.html

- ("Windows kit with embedded Java" or "Mac kit"

If you accidentally close your OpenRefine window, you can type these addresses into the address bar of your web browser:
http://127.0.0.1:3333/
http://localhost:3333
Note, you have to have the OpenRefine service running for these links to work. Otherwise, you'll just get an error.

Dataset: https://raw.githubusercontent.com/LibraryCarpentry/lc-open-refine/gh-pages/data/doaj-article-sample.csv
You can right click on this link in the Etherpad and choose "Save target as" or "Save link as", which will let you save it as a file.
You can also use the Web Addresses (URL) option and paste the above link in. OpenRefine will download the data directly from the GitHub site.

How many facets do you have in the Publisher field?
7
7
7
7
7
7 "choices"
7
7
7
7
7
7
7
seven
8 b/c I still had the first row of subjects in the columns when I opened this :)
You may want to re-import the file and specify that your file has headers. Otherwise it could get confusing later!
Can I delete that row? I renamed all of the headings of the columns
As long as the headers match, I think that should be okay!
? How do I delete that row? :)

You want to either facet or filter so that only that row is visible. Then, under the All column, you will see an option to Edit Rows > Remove matching rows. This will remove all the rows currently visible. In general, it might be a little risky to do corrections like this - in the future, I would recommend reimporting the file and starting over fresh, as long as you aren't too far in the cleaning process!
SUCCESS, graz
Glad that worked for you!

Break: Return at 10:08
Even only seeing the split/join, facet, and filter features so far, do you have an idea for a project you might use OpenRefine? What is it?

Cleaning and making a new Appendix for pigment data
I get data for our serials/databases/ebook usage. Cleaning up is a big chore
I have similary citation information that would love to ve cleaned up easily like this.
Finding aids--especially container lists
Making sure rights information is uniform
When I send reports to people, there are some cleanups that would probably be easier to do in OpenRefine than in Excel
Cleaning up index terms
I'm already using OpenRefine for projects with FAST headings and WikiData. Very useful for clean-up!
Finding text fields with numbers, punctuation, etc. in them. Filtering with regular expressions is the best!
Finding aids - subject headings

Reference for GREL functions: https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions
Using functions is super powerful for cleaning data - I highly recommend looking into them!
(cheat sheet of some common needs: https://guides.library.illinois.edu/openrefine/grel)

To find authors in the Lastname, Firstname format and change them to Firstname Lastname format instead:

```
value.match(/(.*),(.*)/).reverse().join(" ")
```

Formulate the URLs to fetch from Crossref

```
"https://api.crossref.org/journals/" + value
```

Title: value.parseJson().message.title

Retrieve data from any REST API: Edit column > Add column by fetching URLs
Retrieve data from dedicated reconciliation services: Reconcile > Start reconciling

Q: what was the significance between that publisher one or two checkmarks? the matching?
A: Clicking one checkmark chooses the value for just that record. Clicking two checkmarks changes that value for all of those entries in the whole dataset.

URL for VIAF reconciliation service: http://refine.codefork.com/reconcile/viaf

Add VIAF ID column Pulls the Identifier (can be used for Wikidata as well): cell.recon.match.id

More details on reconciliation services: https://librarycarpentry.org/lc-open-refine/13-looking-up-data/

Create your own reconciliation service from a CSV file: https://github.com/amdevine/openrefine-reconcile-csv. Warning, this is a little finicky and may require some experimentation.

Carpentries Brown Bag: using Python scripting in OpenRefine
https://github.com/SmithsonianWorkshops/carpentries-brown-bag/tree/master/2019-08-22-scripting-openrefine

I have used this to great effect when I had a lot of values I wanted to translate and I didn't want to write a ton of nested if/else statements.
I've also used python in OpenRefine to great success, it can take a little fiddling to get the syntax right, but was worth all the effort.

Daily Feedback Survey: https://docs.google.com/forms/d/e/1FAIpQLSckvNc-MCKaW11B55YYAYhPS-OYRawA8RIJugDeG4VoyUX6mw/viewform

----

*Day 4: Introduction to Git

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

Git lesson: https://librarycarpentry.org/lc-git/

***Attendance Day 4**

Please add your name and your favorite thing you learned about OpenRefine
Allie Alvis-amazing data-understanding program!
Carrie Smith - so much cleaning -- the dates, the VIAF ref, merging clusters
Heidy Berthoud - need more practice with GREL
Daniel Euphrat figured out how to use Openrefine to deduplicate
Alison Oswald-great program
Stephanie Kurasz
Eden Orelove - query APIs, esp Wikidata
Suzanne Pilsk: More more more - Open Refine is great. I would like to have applied more of our RegEx skills in Open Refine - And the importing and exporting results - but I see so much possibility and have already got plans.
Sue Graves - OMG - the possibilities for cleaning
Mike Trizna: I was fascinated by the difference between records and rows
Manuel Samudio - favarite thing about openrefine is creating new colums with regex
David Holbert - such a robust program is free!
Bess Missell - I'll be using OpenRefine more than Excel now!
Amanda Devine - I learned how to query APIs from OpenRefine!
Mitch Toda - OpenRefine looks like a powerful way to clean up data and much easier to work in than Excel
Daniela Jiménez - being able to use APIs in OpenRefine
Stefaan Hurts - a lot to take in at once but I'm sure it'll have its uses in the future
Bonnie Felts - I could only make the first hour, but I had never seen OpenRefine, so just the basics of it was really cool.
Richard Naples - getting data from somewhere out there using get from url and resolver

Git is language of the remote repository - source code management
GitHub is a website - a destination - distributed: working copy on your pc, and local repository - then you push and pull to the server (remote repository). Update/Commit local Pull/Push to the remote server
Fork - take a project/document and make your own changes. Not making changes to the original source

On Windows, Git BASH defaults to my default user directory.
pwd
^- prints out my current working directory, /c/Users/Amanda
I can switch to my Documents folder with
cd /c/Users/Amanda/Documents
If I make a folder called Carpentries in my Documents folder, I can switch to that with
cd /c/Users/Amanda/Documents/Carpentries

To exit the vim window, press the Esc key, then type :wq

Break until 10:45 EDT

Questions for breakout groups:
In groups:

- illustrate the concepts discussed in the first hour
- try to 'draw' what different commands mean
- try to come up with synonyms for what the commands are doing.

Example text for html:

```
<!DOCTYPE html>
<html>
        <head><title> My GH Page </title></head>
        <body>
                <h1> Hello World </h1>
        </body>
</html>
```

Git cheat sheet: https://education.github.com/git-cheat-sheet-education.pdf

You can also create web pages with GitHub Pages from markdown files. Instead of having to type HTML, you can type basically in plain text, with a few special characters to indicate formatting. It's really easy - if you're curious, feel free to send me (Amanda) an email for more details.

For one of our Carpentries Brown Bags, I created a tutorial entirely in an R Notebook (an RStudio file format that incorporates markdown text and R code). It was really easy to share this R Notebook with GitHub pages. https://amdevine.github.io/cbb-r-mapping/

A great R-specific resource for using Git: https://happygitwithr.com/

Wrap-up: https://smithsonianworkshops.github.io/si-carpentries-handbook/workshop_outro.html

Daily Feedback Survey: https://docs.google.com/forms/d/e/1FAIpQLSckvNc-MCKaW11B55YYAYhPS-OYRawA8RIJugDeG4VoyUX6mw/viewform