Spreadsheets and OpenRefine

Adela Sobotkova

History and Classical Studies

29th August 2022

Links to Instructions

https://datacarpentry.org/spreadsheets-socialsci/setup.html

https://datacarpentry.org/openrefine-socialsci/setup.html

https://librarycarpentry.org/lc-open-refine/13-looking-up-data/index.html

https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine

https://datacarpentry.org/r-socialsci/setup.html

Introduction to Spreadsheets

Guiding question for this episode:
What are basic principles for using spreadsheets for good data organisation?

Introduction to Spreadsheets - Exercise

In GDrive shared document (`http://bit.ly/DigitalMethodsReflections`):

- How many people have used spreadsheets in their research?
- How many people have accidentally done something that made them frustrated or sad?

Formatting data tables in Spreadsheets

livestock_owned_and_numbers
1, (poultry)
3, (oxen , cows)
1, (goats)
4, (oxen , cows)
10, (oxen , cows , goats , poultry)
1, (goats)
1, (oxen)
2, (oxen , goats)
3, (oxen , goats)

Figure 1: Data Carpentry: combined info (CC-BY)

poultry	cows	goats	oxen
1	0	0	0
0	2	0	1
0	0	1	0
0	3	0	1
5	2	2	1
0	0	1	0
0	0	0	1
0	0	1	1
0	0	2	1

Figure 2: Data Carpentry: single info (CC-BY)

Exercise

We're going to take a messy version of the SAFI data and describe how we would clean it up.

- Download the messy data.
- Open up the data in a spreadsheet program.
- Notice that there are two tabs. Two researchers conducted the interviews, one in Mozambique and the other in Tanzania. They both structured their data tables in a different way. Now, you're the person in charge of this project and you want to be able to start analyzing the data.
- With the person next to you, identify what is wrong with this spreadsheet. Discuss the steps you would need to take to clean up the two tabs, and to put them all together in one spreadsheet.
- Document your group's thoughts in the GDrive shared document.

Important Do not forget our first piece of advice, to create a new file (or tab) for the cleaned data, never modify your original (raw) data.

After you go through this exercise, we'll discuss as a group what was wrong with this data and how you would fix it.

Metadata

"Data about data"
Exercise (maybe):

- Download a clean version of this dataset and open the file with your spreadsheet program. This data has many more variables that were not included in the messy spreadsheet and is formatted according to tidy data principles.
- Discuss this data with a partner and make a list of some of the types of metadata that should be recorded about this dataset. It may be helpful to start by asking yourself, "What is not immediately obvious to me about this data? What questions would I need to know the answers to in order to analyze and interpret this data?"

Dates in Spreadsheets

... are stored as integers.
Excel stores dates as numbers - see the last column in the above figure. This serial number represents the number of days from December 31, 1899. In the example, July 2, 2014 is stored as the serial number 41822.

	A	B	C	D	E	F	G	H	I
1	What I typed in	day-month	DOW, month, day, year	month-year	Initial-year	M/D/YYYY	DD/MM/YYYY	DD/MM/YY	number
2	2-jul	2-Jul	Wednesday, July 02, 2014	Jul-14	J-14	7/2/2014	02/07/2014	07/02/14	41822
3	Jul-14	14-Jul	Monday, July 14, 2014	Jul-14	J-14	7/14/2014	14/07/2014	07/14/14	41834
4	1-jan-1900	1-Jan	Sunday, January 01, 1900	Jan-00	J-00	1/1/1900	01/01/1900	01/01/00	1

Figure 3: Data Carpentry: combined info (CC-BY)

Quality Assurance - Validation

Use data validation to prevent accidentally entering invalid data.
Consistency and completeness of data in spreadsheets can be vastly improved through validation.
Example: Mounds Table

Exporting Data

Data stored in common spreadsheet formats will often not be read correctly into data analysis software, introducing errors into your data.

Exporting data from spreadsheets to formats like CSV or TSV puts it in a format that can be used consistently by most programs.

Fun Exercise on Danish Kings

Encode the length of rule of individual kings that ruled the Danish territories from the dawn of time to review trends in length of rule over time.
You may use these sites as a source:
`https://kongehuset.dk/monarkiet-i-danmark/kongerakken`
`https://danmarkshistorien.dk/perioder/vikingetiden-ca-800-1050/`

- Document unknowns appropriately
- Calculate the duration of the rule and the midyear of rule from the length of rule dates
- What is the average duration of royal rule across the millennium?
- How many kings rule more/less than the average?

Eventually we will make a scatterplot and linear regression in R to see the long-term trend in Denmark. Repeat for other Scandinavian/European countries if you can!

Breaktime - 10 min

Breaktime, hurray!

Links to Instructions

https://datacarpentry.org/spreadsheets-socialsci/setup.html

https://datacarpentry.org/openrefine-socialsci/setup.html

https://librarycarpentry.org/lc-open-refine/13-looking-up-data/index.html

https://programminghistorian.org/en/lessons/fetch-and-parse-data-with-openrefine

https://datacarpentry.org/r-socialsci/setup.html

Introduction to Open Refine: https://127.0.0.1:3333

"Powerful tool to clean up messy data"

The aspects of OpenRefine we'll breeze through:

- Facets (text, numeric, timeline, scatterplot, custom)
- Filters (you can use a regex)
- Clustering (see below)
- Sorting (exactly what you think, but must fix datatype first)
- Data manipulation with GREL

Introduction to Open Refine:Clustering

Clustering helps you streamline your content auto- magically
`https://docs.openrefine.org/next/technical-reference/clustering-in-depth`

- Key Collision Methods
  - Fingerprint: good place to start, few false positives
  - n-gram fingerprint
  - phonetic fingerprint: useful to spot misspeling or misunderstanding
- Nearest Neighbour Methods
  - Levenshtein Distance: "edit distance", generally applicable
  - PPM: lots of false positives, "last resort"

GREL: Transforming Data in Open Refine

GREL is a General Refine Expression Language, it supports regular expressions! :)
`https://docs.openrefine.org/manual/grel`
Common uses of transformations (GREL) include

- Split data from one column to multiple (addresses, etc.)
- Standardizing the format of data (remove punctuation, lower case)
- Extracting a particular type of data from a string (finding ISBN)

Other Essentials of Open Refine

Scripts and Versions

Exporting data or projects

Advanced Functionality

...

Can you see yourself using Open Refine? Who would you recommend this tool to and why?

Using API with Open Refine

Fetching and Parsing Data - Sonnets - from the Web with OpenRefine

Fetching and Reconciling journal data with OpenRefine

Feedback time

On your green sticky, write one thing I did well today.
On your red sticky, write one thing I could improve upon for next time. Be specific.

Thank you for your hard work today!

Tomorrow 'might' be easier as we'll stick to one tool alone.

Homework

(1) Check Brightspace for HW instructions (OpenRefine Exercises).
(2) Check out the Advanced functions in Open Refine if interested in API.

For Tomorrow and Thursday

(1) Install R and RStudio (you should not pay for any of these)
(2) paste your username on #github channel on Slack. Connect to the Digital-Methods-HASS repository when you get the email invite.
(3) For Thursday: Install git bash (see Syllabus for links) Read the shell / git tutorials.
(4) Start thinking about your final project.