**Roslyn Walker, Student number 45149631**

Learning Journal

FOAR705 Digital Humanities

Week 7 and 8.

Due on 4 October 2019. Learning Journal, Weeks 7-8, Proof of Concept work and Data Carpentry Open Refine, Due week 8 - Part 1

ASSESSMENT CRITERIA

* Acquire core technological skills needed to produce transparent and reproducible research.
* Identify existing digital tools and approaches and apply them to your research.

=======================================================

Contents

[(1) Technology deployment--testing 1](#_Toc21018792)

[(2) Data carpentry exercise notes 3](#_Toc21018793)

======================================================

## (1) Technology deployment--testing

**Objective: Setup example for testing proof of concept**

Steps:

* Establish test summary
* Preconditions
* Datasets Required to test
* Example below

|  |  |  |  |
| --- | --- | --- | --- |
| Test steps: | Result | Expected results | Comment |
| 1. Login | login successful |  |  |
| 1. Click on xxx | go to page |  |  |
| 1. Enter module name, sampling loads | module created |  |  |

Objective: Explore research methods web scrapping

Research on Methodology: Article on source data test: Bay, M. (2018). Weaponizing the haters: The Last Jedi and the strategic politicization of pop culture through social media manipulation. First Monday, 23(11)

Notes from Consultation with Brian 13 September 2019.

**Objective:** Develop an action plan for next steps to Minimum Viable Product Development

**Steps:**

(1) Action data carpentry lessons all way to the end.

(2) Get to R.

(3) Learn how to use R with js files.

(4) Use on own source files.

(5) Expand for use on twitter dataset for the 2016 elections.

By end of break use twitter dataset in R (search google to see if others have accomplished this).

**Resources:**

* Explore application and links from Brian to be helpful to project development and testing.
* <https://datacarpentry.org/r-socialsci/02-starting-with-data/index.html>
* Lessons in data carpentry all to the end <https://cran.r-project.org/web/packages/jsonlite/vignettes/json-aaquickstart.html>
* <https://datacarpentry.org/lessons/#social-science-curriculum> See especially, Data Analysis and Visualization with R for Social Scientists link on the above page.
* <https://stackoverflow.com/questions/16947643/getting-imported-json-data-into-a-data-frame/37739735>
* Use own twitter files as source files, then apply learning to data set at: <https://about.twitter.com/en_us/values/elections-integrity.html#data>
* Investigate functions and operation of the Twitter power track - search Power Track on Twitter.

**30 September 2019**

Objective: Understand how a project brief is formed. Acquire core technological skills needed to produce transparent and reproducible research. Design and develop a digital tool to support your masters research.

Steps:

(1) Proof-of-concept technology deployment

* Design: Scoping Exercise (wk 2, 3)
* Design: Elaboration Plan (wk 4)
* Implementation: Elaboration Results (wk 6)
* Design: Proof of concept design (wk 7)
* Implementation: Proof of concept due (wk 7)

Result: Set of documentation tracking the use of core technological skills to produce transparent and reproducible research. Using new skills to design and develop a digital tool to support masters research.

## (2) Data carpentry exercise notes

Week 7 exercises: Data Carpentry [OpenRefine for Social Science Data](https://datacarpentry.org/openrefine-socialsci/), 13 September 2019

Data Cleaning with OpenRefine. Examining Numbers, Using Scripts, Exporting, Other Resources: <https://datacarpentry.org/openrefine-socialsci/>

>>>>>>>>>

Open Refine notes from class exercise: 13 September 2019

Objective Lesson 4. Examining numbers <https://datacarpentry.org/openrefine-socialsci/04-numbers/index.html>

Steps: To transform cells in the years\_farm column to numbers, click the down arrow for that column, then Edit cells > Common transforms… > To number. You will notice the years\_farm values change from left-justified to right-justified, and black to green in color.

Transform three columns, no\_members, yrs\_liv, and buildings\_in\_compound, from text to numbers.

Result: only numerals (0-9) can be transformed to numbers.

Objective Lesson 5. Using scripts <https://datacarpentry.org/openrefine-socialsci/05-scripts/index.html>

Steps: document the data-cleaning steps applied to the data

(1) In the Undo / Redo section, click Extract..., and select the steps that you want to apply to other datasets by clicking the check boxes.

(2) Copy the code from the right hand panel and paste it into a text editor (like NotePad on Windows or TextEdit on Mac). Make sure it saves as a plain text file. In TextEdit, do this by selecting Format > Make plain text and save the file as a .txt file.

Result: All changes are tracked in OpenRefine, this information can be used for scripts for reproducing an analysis.

Script saved to Github as text file.

Objective Lesson 6. Exporting and saving data from OpenRefine

Objective Lesson 7. Other resources.

* changing how many rows you view doesn’t change how many are selected
* go to view under a column, can collapse columns in various ways
* help us with interacting with online databases? - they use the same ideas and terminology e.g. facet searching on Trove

2:46

edit two things - text (abc) and 0

apply a numeric facet

uncheck numeric (but leave non-numeric and blank checked) - that shows me the blank column and the one with text, fix them (no in facet no text boxes)

3:55? check when I saved the .txt file on desktop

in Undo/Redo

click extract

produces script in right panel, copy this

paste into a text editor

in notepad - go to language, json highlights the things in different colours

then when I create a new project with the original data (hasn’t been changed in OpenRefine), call it something different

go to undo/redo

click apply

paste the copied text from the text file

and then click perform

export project if I want to move it to another computer etc.

can put that onto GitHub

>>>>>>>>>>>>>>>

Class catch up. 1 October 2019. Learning Journal submission, 2pm. 4 October 2019

Week 6 exercises: Data Carpentry--Data Cleaning with OpenRefine - https://datacarpentry.org/openrefine-socialsci/

Lesson on Introduction, Working with Open Refine, Filtering and Sorting

**Objective: Setup for lesson on OpenRefine**

Steps: download <http://openrefine.org/> for Mac.

* Chrome browser is installed and set as default browser.
* OpenRefine runs in your default browser. It will not run correctly in Internet Explorer.
* Download software from [http://openrefine.org](http://openrefine.org/) -- OpenRefine 3.2
* Unzip the downloaded file into a directory by double-clicking it. Name that directory something like OpenRefine.
* Go to your newly created OpenRefine directory.
* Launch OpenRefine
* Drag icon into Applications folder, and Ctrl-click/Open it.
* Note: If using a different browser, or OpenRefine does not automatically open for you, point your browser at http://127.0.0.1:3333/ or http://localhost:3333 to launch the program.

Error message: “OpenRefine” can’t be opened because it is from an unidentified developer.

Solution: <https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions>

macOS: If you get a message saying "Open Refine can't be opened because it is from an unidentified developer"

* open System Preferences and go to
* "Security and Privacy" and the
* General tab. Here you will see a message indicating that "OpenRefine was blocked from opening because it is not from an identified developer".
* Click the "Open Anyway" button to complete the OpenRefine installation.

**Objective. Complete lesson on Introduction to OpenRefine for Social Science Data. What is OpenRefine useful for?** <https://datacarpentry.org/openrefine-socialsci/01-introduction/index.html>

Teaching: 10 mins.

Noted: Background to application. An open source tool to clean up messy data and transform its’ format. OpenRefine cleans and formats data and tracks the changes made. Data workflow is used in preparing the data for analysis such as data cleaning. Data cleaning is where errors are found and corrected, and the format can be made consistent. This step forms reproducibility as the analysis. Differentiate data cleaning from data organization.

Note use and application of OpenRefine.

* actions can be reversed in OpenRefine.
* when using save it becomes a new file--OpenRefine copies the data and does not modify the original dataset.
* data cleaning steps often need repeating with multiple files. OpenRefine keeps track of all of actions taken and allows them to be applied to different datasets. Save as txt file.
* concepts such as clustering algorithms are quite complex, but OpenRefine makes it easy to introduce them and use them.

**Features of OpenRefine**

* Open source is on GitHub: <https://github.com/OpenRefine/OpenRefine>
* Community of practice ready to help.
* Works with datasets (100,000 rows) and adjusts memory allocation to accommodate larger datasets.

Objective. Complete lesson on Working with Open Refine to OpenRefine for Social Science Data. <https://datacarpentry.org/openrefine-socialsci/02-working-with-openrefine/index.html>

Teaching 15 mins: Exercises 20 mins.

* How can we bring our data into OpenRefine?
* How can we sort and summarize our data?
* How can we find and correct errors in our raw data?

Steps: Create a new OpenRefine project

1. Downloaded data obtained from interviews of farmers in two countries in eastern sub-Saharan Africa (Mozambique and Tanzania). Filename SAFI\_openrefine.csv Open file in OpenRefine.

2. Create new project.

3. Get data from. This computer.

Result: opened SAFI\_openrefine csv

**Objective: Use facets—explore data by applying multiple filters**

Steps for exercise:

Use faceting look for potential errors in data entry in the village column.

1. In the village column. Click the down arrow and choose Facet > Text facet.
2. sorting this facet by name and by count.
3. Hover the mouse over names for edit function
4. fix errors:

* Change Chirdozo to Chirodzo.
* Ruca to Ruaca.
* Correct Ruaca - Nhamuenda and Ruaca-Nhamuenda
* 49  an error fix unsure

Exercise:

1. interview\_date values = 19 unique

2. change format through edit cells> common transforms>to date

3. Face> timeline facet

Most of the data was collected in November of 2016.

**Objective using clustering to detect errors**

Steps:

* In the village Text Facet we created in the step above, click the Cluster button.
* In the resulting pop-up window, you can change the Method and the Keying Function. Try different combinations to see what different mergers of values are suggested.
* Select the key collision method and metaphone3 keying function. It should identify two clusters.
* Click the Merge? box beside each cluster, then click Merge Selected and Recluster to apply the corrections to the dataset.
* Try selecting different Methods and Keying Functions again, to see what new merges are suggested.
* You should find no more clusters are found. None of the available methods offered to cluster Ruaca-Nhamuenda with Ruaca or Chirdozo with Chirodzo. To merge these values we need to hover over them in the village text facet, select edit, and manually change the names.

Results:

* Changed Chirdozo to Chirodzo and Ruaca-Nhamuenda to Ruaca. You should now have four clusters: Chirodzo, God, Ruaca and 49.

**Objective: Lesson Transforming data.**

Steps:

Click the down arrow at the top of the items\_owned column. Choose Edit Cells >

This opens a window, type a GREL expression. GREL stands for General Refine Expression Language. First we will remove all of the left square brackets ([). In the Expression box type value.replace("[", "") and click OK. What the expression means is this: Take the value in each cell in the selected column and replace all of the “[” with “” (i.e. nothing - delete).

Click OK. See in the items\_owned column there are no left square brackets.

Repeat to remove the single quote marks ('), the right square brackets (]), and spaces from the items\_owned column.

Result

* value.replace("'", "")
* value.replace("]", "")
* value.replace(" ", "") You should now have a list of items separated by semi-colons (;).

Steps:

* Click the down arrow at the top of the items\_owned column. Choose Facet > Custom text facet...
* In the Expression box, type value.split(";").

Click OK.

Result: most commonly owned are mobile phone and radio, least commonly owned are cars and computers.

Exercise: Clean up steps and customized text faceting for the months\_lack\_food column. Which month(s) were farmers more likely to lack food?

Result: All four cleaning steps can be performed by combining .replace statements. The command is: value.replace("[", "").replace("]", "").replace(" ", "").replace("'", "") This can also be done in four separate steps if preferred. November was the most common month for respondents to lack food.