**Roslyn Walker, Student number 45149631**

Learning Journal

FOAR705 Digital Humanities

Week 4. 22 August 2019

Task one: Data carpentry

Read and do all exercises. Exercises should be recorded in your Learning Journal and uploaded to cloudstor.

1. <https://datacarpentry.org/spreadsheets-socialsci/00-intro/index.html>
2. <https://datacarpentry.org/spreadsheets-socialsci/01-format-data/index.html>
3. https://datacarpentry.org/spreadsheets-socialsci/02-common-mistakes/index.html

*Task note: For learning journal add 2 examples of problems in data produced by your discipline. You will get a “Well done indeed” from me if you can find these problems in published datasets in your discipline, and cleaning one of your disciplines’ datasets could very well be your Proof of Concept.*

**Task one: Data carpentry exercise.**

Objective: Complete learning task <https://datacarpentry.org/spreadsheets-socialsci/00-intro/index.html>

Action: 15 min duration with exercises 3 min. Open Microsoft Excel.

Noted text of problems with spreadsheets. The graphical interface can complicate replicating steps. Data quality can be compromised by errors introduced through applying formula to neighbouring cells.

*Starting exercise: How many people have used spreadsheets in their research?*

*How many people have accidentally done something that made them frustrated or sad?*

I answered these questions in the class sheet on Github. No major problems encountered with spreadsheets so far.

Result: Start. 22:14 Completed. 22:19

**Task two: Data carpentry exercise: formatting data tables in spreadsheets.**

Action: 15 min duration, 15 min exercises. Start 22:21. Complete 22:47

Noted importance of using well-formatted tables from the outset. Note best to automate conversion for optimal layout/format where different software or interface requires it.

1. Open messy data sheet.

Note: Messy has two tabs. The data table formats are inconsistent.

the spreadsheet has three tables. Mixed use of numeric and text which should be consistent. Difficult to read both tables because of inconsistent formatting. Information/data is confusing. Zeros should be added to record zero. Missing data should have null entry. Comments should be in a separate column not mixed in the data columns. The units of the measures should appear in the header for each column. Special characters should not be used in data headings or in the data.

2. Open clean version of data.

Note: 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? What types of metadata that should be recorded about this dataset.

It is not clear what the questions were for the data results, this could be noted. Metadata is free text for comments, units used in this sheet they disrupt the format of the data file and better placed in a codebook. Use a standard, such as Data Documentation Initiative (DDI) for reference: <http://www.ddialliance.org/>

**Task three: Data carpentry exercise: formatting problems** <https://datacarpentry.org/spreadsheets-socialsci/02-common-mistakes/>

Action: 20 Min duration. Start: 22.51 End 22:59

Note: Observe errors in tables.

Adding here for ready reference, tidy data rules noted as key points to this section:

* one table for one spreadsheet.
* keep data in one tabs
* use a new tab and copy when cleaning the data up
* note zeros as zeros [0]
* use null value for missing data.
* avoid formatting for presentation
* keep comments in a separate column
* place unit of data in the column header
* place one piece of information in each cell, avoid multiple on one cell
* don’t use special characters in the data
* don’t use spaces, numbers or special characters in the column header

Date: 15 August 2019. 20:52

**Example one: Download Twitter Content.**

**Objective: Download data from my Twitter account.**

As a test example of data retrieval for my discipline: Journalism. I am downloading data from my own Twitter archive. Following instructions posted at: download your Twitter archive: https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive

**Action from Twitter help menu:**

1. Go to your Account settings by clicking on the more icon in the navigation bar, and selecting Settings and privacy from the menu.
2. Under the Account section, click Your Twitter data.
3. Enter your password under Download your Twitter data, then click Confirm.
4. Click the Request data button. If your Twitter account is connected to Periscope, you’ll have the option to Request data from Periscope as well.
5. When your download is ready, we'll send an email to your connected email account. From your settings, you can click the Download data button under the Download data section.

**Error:**

Needed to reset password to account, as password is saved and hidden. Completed.

1. Step one, there is no more icon, click on the profile picture, Settings and privacy is in the drop down menu.

**Result:**

Twitter advises in a note on help that it may take a few days for them to prepare the download of my Twitter archive. I hadn’t received the data by the end of this training session. I will check again tomorrow. This exercise flags data preparation and may take several days to retrieve from Twitter.

I am interested in seeing how they return the data and formatting/presentation whether the data will need to be formatting for consistency and useability in a spreadsheet format.

**Example two: Preparing metadata for digitisation of archives project.**

**Objective:** Set up a data table to collate categories and identifying tags for an online database of digital archives.

I am building a data set of metadata tags for my Research Internship placement. My host organisation started building a list in a word document as tables. There were multiple tables and inconsistent data entry in the tables.

Action: Open new spreadsheet in Excel.

1. Import table from Word to Excel by copying table in Word and pasting it to Excel
2. I’ve opened a new spreadsheet and placed the data in the first tab.
3. In a new tab titles revised tag, started to clean data.
4. Removed formatting and changed heading using checklist for tidy data rules

Result: Placing the data in a spreadsheet provides easier access and viewing of complete set of tags. I can view the gaps in the data. I can clean up descriptors and edit for consistency.

I can build a tag list without fussing with formatting and table building in Word.

16 August 2019.

**Objective:** Open data set received from Twitter.

Action: I received an email yesterday as a result of my inquiry to Twitter to access my data on Twitter.

1. The email was sent to an email account I don’t frequently used, but often link to applications to screen junk mail. The email was received within an hour of the request. The email provides a link to download, that is noted as expiring in a month.
2. The link goes back to my Twitter account and asks for my password. I entered my password and arrived at new page with a download link.
3. Download is a twitter zip file with 47 items, with two folders profile\_media and tweet\_media, files have a .js extension.

Error: Unable to utilise data without further knowledge on which files to access or what data I want to retrieve. Note the term your Twitter data includes range of a files including periscope.

19 August 2019.

**Objective.** Submit assignment in La Tex on Cloudstor.

**Action.** I worked through LaTex and converted my scoping study by pasting text into Overleaf and formatted it through RTF. I exported to PDF okay and submitted, and exported in format as .zip file to submit on Cloudstor.

**Error.** How do I preserve .tex when uploading to Cloudstor? Files in the submission folder appear to be .tex and pdf and zip. I repeated steps and got the same result. Why are some folders in the submission folder bright orange folders and others file extensions as noted. Some tips and steps for the best path would be great