**Roslyn Walker, Student number 45149631**

Learning Journal

FOAR705 Digital Humanities

Week 4. 22 August 2019

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

(1) Technology deployment

(2) Data carpentry exercise notes

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

(1) Technology deployment

23 August 2019: Objective: Understand application of Computational thinking.

Steps: noted with reference in Slack:

Decomposition: Breaking down data, processes, or problems into smaller, manageable parts

Pattern Recognition: Observing patterns, trends, and regularities in data.

Algorithm Design: Developing the step by step instructions for solving this and similar problems

* First (and most important), decompose the activities that involve ‘pains’, opportunities for ‘gains’, and any solutions you proposed into small parts and/or discrete steps.
* Next (if possible), identify patterns in the problems you are trying to solve or the solutions you are proposing.
* Finally, revise the solutions you developed during BA to produce a step-by-step guide describing what you want to accomplish.
* To rephrase the above, scoping II now takes the amorphous pains/gains from Scoping I and will "output" a list of steps, a recipe, an algorithm, for individual tasks which must be done (even if you don't know how to do them) to go from start to finish.

Result: specifically interested in automated transcription for better ways to do reference from Brian to Alveo's transcriber tool: "Data storage with appropriate metadata > Loading it into a program which can extract audio from individual files > Sending that audio to The Cloud > Rendering the returns as appropriate text > Saving the text and metadata into a machine readable format that humans can also use".

Next steps: consider steps in application of automated transcription of social media under hashtag publics and counterpublics. Retain message and image frame.

**23 August 2019: Objective. Task for Scoping II. Computational thinking.**

Action: seeking advice on path and application relating to repository and data organisation for image objects as archives.

Steps:

* Noted in my scoping paper tasks for developing a dataset for cataloguing and adding contextual information about works in an archive collection (i.e. thematic tagging, making connections between objects, people, places).
* Take an existing list in a word document and place this information into a spreadsheet format (Excel).
* The list for word has been generated from a list of terms used by other ephemeral art institutions in their archive repositories. I've noted these steps in my learning journal about starting work on the data set.
* My question in developing this data: What conventions should I use and work on first -- some ideas from today: coding sheet of terms, Readme files, control vocabulary --that will make my task of tagging digital records easier when posting them to the repository.
* How do I start to generate a useful list of terms for tagging an object in the repository?

Result: Advice from Shawn: To answer your question: you're in the realm of a number of existing platforms / systems. I think we discussed this last week: you are somewhere between [https://tropy.org/](https://slack-redir.net/link?url=https%3A%2F%2Ftropy.org%2F) and a content or document management system (CMS or DMS). You could even use bibliography software like Zotero to do this (it's worth investigating during elaboration at least).

Next steps. Trial applications across: bibliographic management: content and document management, and image repositories. Consider specific application to archive of social media content and images in preparation for analysis and curation. Test data thematic tagging and process data collection from source to filter content.

19 August 2019. **Objective.** Submit assignment in LaTex 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.

Next steps: Solution some folders have tex files in them submission as described above is okay.

**19 August. Objective: clarify scoping task for submission.**

Action: I feel I've gone off topic in my scoping exercise by using my PACE research internship role and task to produce a virtual exhibition of documentation through a digitised archive. I've added a note that I am using the proof of concept to build a repository for images online from social media that will form data for my thesis.

This means two tasks: (1) develop proof of concept building the virtual database and (2) application to social media.

I note that instructions asked for "Think about a major research project you have completed, and about your thesis. Imagine yourself going through the process of producing a 100+ page thesis and: "I wonder if this might put the exercise out of sync somewhere and if I should focus on page production instead?

Seeking advice from Brian on approach.

Result: Advice response from Brian: “I think having a specific and productive need for your proof of concept is far better than generic thesis support. One of the things we are grading for in the scoping is "hole in the literature" (for HD, I mean) -- so if you can demonstrate there don't exist well documented solutions for virtual exhibitions, especially as they apply to X, I feel that satisfies criteria well. The way to think about it is more "How would I get a journal article out of this?"

Error. Focus on Scoping paper II and defining the boundaries of what is in and out of the scope of the project: requires definition of boundaries eg what’s in and what’s out. Include the type of data to be used in the proof of concept. On technical aspects, commit in version control and document in Learning Journal.

Next steps. Address errors and consider/explore aspects of “hole in the literature” that would be improved by a technical solution.

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.

Action: Sought advise from Brian, note to try to look at the file in a real text editor like [atom.io](https://slack-redir.net/link?url=http%3A%2F%2Fatom.io) or sublime text. There are ways of reading json, but step 1 is to take a look at it, and to look for the metadata.

Next steps: I will return to trial this approach. I would also like to download my Instagram data

**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.

**16August 2019. Objective: explore version management on Github**

Action: Asked question to channel: When uploading a new version of a document to github, and the file name has changed will version control apply to the document along with earlier versions of the document under different title names?

Result: Advice from Brian. Github tracks changes line by line in your documents, and tries to identify your documents by content. When it's a zip file (i.e docx) then it's much harder for git to see inside. (For a fun time, rename the .docx to .zip and see what's inside).

However, for text based files, yes it will absolutely track their names as they change. There are also manual ways to tell git that it's the same file.

Result: uploaded files to Github and downloaded, next tested new versions on existing documents and annotating in commit.

**(2) Data carpentry exercise notes**

Week 4. 23 August: Spreadsheets for Social Sciences https://datacarpentry.org/spreadsheets-socialsci/

Action: Export the csv. View it in a text editor like Atom.io, Sublime Text, or notepad++ Think about the benefits of an always-readable and not tied to a subscription or specific program data format.

1. Lesson objective: Dates as Data. Do tasks for data organization and practices for effective data wrangling

Action. 1 10 mins. Dates as data. Dates are inconsistent in presentation which is problematic for data automation. Fact: In Excel date is stored a number 41822, and 41822 + 90 = 41912 which Excel interprets as the 30 September 2014. Regional variations of date cause data errors, best to note date data (month, day, and year).

Action. Tasks, split dates into component value for ease in handling. Action for exercise: add three new columns, input month, day, year as numeric (general) entry. Saved file.

Action. 2. Default year. Noted spreadsheet defaults to current year, where no value is applied. Can throw out data through the date default. Note to use caution with historical data. Excel translate post 1900 dates into internal format, resulting in mixed data.

2. Lesson objective: Quality assurance—lessons to limit error in data entry.

Note. Validate data on input. Age should be numeric greater than 0 and less than 120. Can note these as example of acceptable values for cells.

Action: Restricting data to a numeric range and Restricting data to entries from a list.

* Step one. Copy data to new tab. Task using data sheet titled. SAFI\_clean. Selected no\_members column amended **data validation field** for range minimum 1 and maximum 30. Tested result. Working okay.
* Step two. Task add message for error return. **Input Message**.
* Step three. Customised resulting message for years living for a data range between 0 and 120 using **Input Message**.

Action: Restrict data to entries from a list. Followed action using **Data Validation**, Settings, added an input message: grass, muddaub, burntbricks, sunbricks, cement. Drop box for controlled vocabulary now functional.

3. Lesson objective. Exporting data. To practice exporting data from spreadsheets and store in a universal file format.

Step one. Open **file** and **save as** csv. Note: CSV files can be opened in Excel. File saved as SAFI\_clean\_2.csv

Note: When using data that contains commas, enclose the fields with double quotes

**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

**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.

Question: 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.

Result: 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 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

# Learning Journal week 1 action notes:

Note: Scoping Exercise is not due next week, due in week 3 instead.

Objective: Think about frustrating and repetitive practices in your research process from collection to publication. Key concepts and terms for drafting, with reference to key authors.

Result: drafting notes for scoping exercises.

Objective: Try to restore a file from a 6 month or older backup. Document how it goes.

Result: I use the cloud, so I can access my files from multiple devices. It’s worked well so far. Only one glitch two years back when I muddled up passwords for devices and the cloud locked me out. After a day on the phone to Apple genius, recovered cloud and device access okay. I’ve sometimes had trouble with version control between the laptop and desktop computers.

Objective: Sign into <https://foar705.slack.com/> and say hi. Take a look at two year old convos.

Action: account active in Slack. Viewing #random #general channels and direct messaging.

Action: Sign into cloudstor: [https://cloudstor.aarnet.edu.au/plus/index.php](https://slack-redir.net/link?url=https%3A%2F%2Fcloudstor.aarnet.edu.au%2Fplus%2Findex.php) and let us know in slack when you have, so we can add you to a group.

Result: signed into cloudstor okay.

Objective: Look into "project management tools" like Trello, Jira, and Asana, and form opinions.

Result: I’ve used Slack for professional coordination of an online journal, it works well however the organization did not purchase the full use, so files were wiped after seven days. That meant we had to use additional tools to document resource files and links. As this was standard operating procedure it worked okay. However, team members often cited, oops message lost, gobbled by Slack. Also used Slack for coordination of team in a Tedx project.

Objective read: [https://impossiblehq.com/an-unexpected-ass-kicking/](https://slack-redir.net/link?url=https%3A%2F%2Fimpossiblehq.com%2Fan-unexpected-ass-kicking%2F)

Extra optional: Read program or be programmed by Rushkoff

Result: material linked in references with summary entry.

Objective: Sign into: [https://github.com](https://slack-redir.net/link?url=https%3A%2F%2Fgithub.com) and tell me your username on slack.