Basic Data Handling with Excel

An introduction to research data management

Western Libraries

What is Research Data Management?

- The organization and maintenance of data throughout the research process
- Incudes
  - Setting up plans and processes before starting data collection
  - Keeping track of, documenting, and backing up data during the research project
  - Archiving or publishing data after the project has completed
- If you collect or use data you are doing RDM
  - Possibly not *good* (standard, recommended) RDM

Start With a Plan

- Think about what data you are going to collect. Consult with your advisor or collaborators on:
  - Where you will store your data - Network drive? USB? Hard drive?
  - How you will access it
  - How you will keep it secure and backed up
- Particularly important with data on humans, such as survey data.
- Think about formats and analysis. If you are creating data that will be analyzed quantitatively, set it up in rows and columns (variables and cases) using software such as Excel. Each column should have a short descriptive name – names like V1, V2 and so on will make life difficult later on

Documenting and Organizing your data

Western Libraries

Files and folders

- Documentation doesn’t need to mean writing a book about your research project
- A sensible basic documentation setup:
  - Set up a sensible folder structure
  - Use intelligible file names
  - Give columns of data in your file sensible headers
  - Have a readme file for each folder (or in some cases, each data file)

Documentation

- Settle on a file naming convention. “Descriptive_name-change_made-date.xlsx” often works.
  - Bad: “FinalData.xlsx”, “ThisIsReallyTheFinalVersion.xlsx” and “FinalFinalDataFIXED.xlsx”!
  - Better: “CodedTweets-random_subsample-Feb3_2021.xlsx”
- Give columns intelligible names
- The main principle in choosing names, whether for variables / column headers or for files, is *short, but meaningful*.
- Make sure you note down every part of how you collect your data as you are doing it – dates, decisions made, choices to include or exclude. These can be kept in a readme file
- If you are categorizing things keep a detailed explanation of which codes correspond to what!

Common information for a readme file assuming one collection of related data per folder

- Title, short description for the data collection
- Person responsible for collecting the data (presumably you)
- For each data file in the collection (or subcollection in case of e.g. sets of image files)
  - Short description of what data it contains
  - Date(s) of collection and download
  - Methods of data collection
  - Describe any data processing and indicate responsible person
  - Missing data codes
- Adapted from Cornell University, Guide to readme style metadata

Simple Readme for Kristi's Twitter Research

- Title: Tweeting is Open
- Description: Research into how open data hashtags are used on Twitter. Data consists of all tweets in specified time period using the #opendata hashtag. Data includes text of tweet, date, time, hashtags, retweets, tweet author.
- Files:
  - Opendata-twitter-Feb-19.csv – Tweets during the month of February 2019, scraped using R on 1 March 2019
  - Opendata-twitter-Mar-19.csv – same, March 2019, scraped April 1
  - Opendata-twitter-Apr-19.csv – same, April 2019 scraped May 2
  - OpendataTwitterScraper-1-3-19.R – R program used to scrape above, based on Wiley's TwitterscrapR.R script downloaded from http://someURL on 7 November 2018

Well-documented methodology

- Ideally, you should have sufficient documentation on your data that a random stranger who is knowledgeable in your field would be able to
  1. Follow and understand the steps you took to collect your data in the first place and the decisions you made along the way
  2. Take your original data file and reproduce the changes you made to it to get your data into its final form
  3. Reproduce any tables or charts that support your research conclusions
- This will also help if you need to start over, or redo something…

“Reproducibility is collaboration with people you don’t know, including yourself next week.” (Stark, 2014)

Data in spreadsheets

Formats

- Open source or widely standard formats are preferred
- Excel creates particular problems for preservation because of the amounts of hidden information (especially formulas)
- You can export from Excel to text formats like csv, but note it will strip away this information
- Keep a clean copy of your original data and document all the changes and formulas

Some guidelines for using spreadsheets

- Put just one thing in a cell
- Organize the data as a single rectangle (with subjects / cases as rows and variables / features as columns, and with a single header row)
- Column headers should be brief and descriptive
- Create a data dictionary – a separate document explaining what is in your rows and columns
- Do not include calculations in the original data files
- Do not use font color or highlighting as data

- I'm serious about the rectangle

See: Data Organization in Spreadsheets

A rectangular spreadsheet

Idnr	Filename	SentenceBNC	Verb	Particle	Object	Pattern	Object_forn
2	A0D	But by the time he had gathered up her handbag and Lord Woodleigh's camera, which had come to rest nearby, she was able slowly to make her way with them	gather	up	Yes	VPn	NP
3	A0D	Then as soon as the last act goes up Bobby will ring for a doctor and say that Bunty's had an accident.	go	up	No	VP	
4	A0N	James Menzies had locked up his warehouse for the day and come over in time to be included in the lengthening list.	lock	up	Yes	VPn	NP
5	A0X	Owners of small block planes or even smoothing planes can make these planes almost as jointers or try-planes. True up a piece of hardwood (one surface/two sides) the	true	up	Yes	VPn	NP
6	A19	The powerhouse responsible for heating up the dust is hidden from view but the team believes it is either a quasar or a massive burst of star formation.	heat	up	Yes	VPn	NP
7	A19	The only toroidal wound coil is the excitation coil and it should be made up using 0.5mm diameter enamelled wire.	make	up			
8	A6J	'No,' cried Maggie, 'and Mummy, I've messed up your new trousers.'	mess	up	Yes	VPn	NP
9	A6Y	Tony Mason, who has weighed up both the composition and the behaviour of Victorian and Edwardian football crowds as carefully as the sketchy evidence permits,	weigh	up	Yes	VPn	NP

From Digital Humanities Workbench, Structured Data

Characteristics of a good rectangle

- Each row is a case – a single instance of the thing you're examining
- Figuring out what your case is – the fundamental, smallest unit or element of what you are studying – is an important part of constructing your data model
  - Use multiple spreadsheets if you are examining distinct, unrelated things with distinct, unlike features
- Each column is a feature or characteristic of the case – a variable
- Setting up your data this way lets you make use of the features of your spreadsheet program – you can sort on characteristics, count them, and ask questions about the characteristics of your cases like "what percentage of X is Y"
- If applicable, the source of a data item or characteristic should be included as a column

Two rectangular spreadsheets… one is better

Text	Reference 1	Category1	Ref2	Cat2
Iliad	3.1704	Rain	6.832	Wind
Odyssey	7.534	Heat	19.43	Clouds

Needing to repeat characteristics across multiple columns may be a sign you need to rethink your model

This is a better model because the fundamental unit being examined in this model is the reference to weather within the text, not the text as a whole

Reference	Text	Category
3.1704	Iliad	Rain
6.832	Iliad	Wind
11.417	Iliad	Rain
7.534	Odyssey	Heat

Two rectangular spreadsheets… one is better

Text	Reference 1	Category1	Ref2	Cat2
Iliad	3.1704	Rain	6.832	Wind
Odyssey	7.534	Heat	19.43	Clouds

Needing to repeat characteristics across multiple columns may be a sign you need to rethink your model

This is not a real rectangle. If you need to sort and analyze data by category, you lose track of which text each reference is in. Many features of Excel won't work with merged cells.

Reference	Text	Category
3.1704	Iliad	Rain
6.832		Wind
11.417		Rain
7.534	Odyssey	Heat

Western Libraries

Two versions of the same spreadsheet

Year	Film Name	Category
1976		
1977	a muerte de Sebastián Arache y su pobre entier	Official Selection
	Barra pesada	Official Selection
	La casta divina	Official Selection
	Rodin mis en vie	Short Films
	Doña Flor e seus dois maridos	Special Events
1978	O desconhecido	Official Competition
	Lovizna	Official Competition
	Noitada de samba	Official Competition
	Chuquiago	Latin American Cinema
	Daniel, o capanga de Deus	Latin American Cinema
	O jogo da vida	Latin American Cinema
	Parada 88	
	Gamin	
	Cantata de Chile	
	Cascabel	
1979	La isla	
	Hasta cuando..?	
	Amor bandido	
	Samba da criação do mundo	

Year	Film Name	Director	Category
1977	La muerte de Sel	Nicolás Sa	Official Selection
1977	Barra pesada	Reginaldo	Official Selection
1977	La casta divina	Julián Past	Official Selection
1977	Rodin mis en vie	Alfred Brai	Short Films
1977	Doña Flor e seus	Bruno Barr	Special Events
1978	O desconhecido	Ruy Santos	Official Competition
1978	Lovizna	Sergio Olh	Official Competition
1978	Noitada de samb	Carlos Tou	Official Competition
1978	Chuquiago	Antonio Eg	Latin American Cine

Only entering the year once may save time or look cleaner… but if you need to sort or filter your data on some other characteristic you're in trouble. First sheet was pretty but unusable.

Western Libraries

Scary data bedtime stories

- Bumped USB key sticking out of a computer, it snapped. Didn't back it up. (Student.)
- Excel file got somehow corrupted and can't be opened. Didn't back it up, didn't have a clean copy from before changes were made… (Colleague.)
- Hired a student to collect and organize the data, they graduated, now has no idea what any of these files are (Researcher who asked me for help.)
- Set up a spreadsheet with columns Q1, Q2 and so on, with a document explaining everything. Then a question got added and the document didn't get updated…
  - There's probably a data librarian purgatory where the documentation all has notes saying that V17 is actually Q16 in the table on page 57, and V18 in the most recent version of the file, and nothing is dated so you can't tell if most recent means finaldata.csv or fixeddata.csv

Further reading and support

- Data Organization in Spreadsheets
- Digital Humanities Workbench, particularly
  - Structured data
  - Data Modelling
- Data management for the Humanities

For more information or to schedule an appointment, contact:

- Kristi Thompson, kthom67@uwo.ca
- Our joint team address: rsclib@uwo.ca (I will respond to messages sent here, or a colleague if I am out)

Western Libraries