Data Organization

Preservation and Curation of ETD Research Data and Complex Digital Objects

INSTITUTE of Museum and Library SERVICES

EDUCOPIA INSTITUTE

Workshop Background

Purpose

- Provide you with resources and tools to help you address the challenges and opportunities “data organization” methods pose and provide for you as a researcher.

Context

- Workshop Series: Preservation and Curation of ETD Research Data and Complex Digital Objects
- Other topics: Copyright, File Formats, Metadata, Storage, Version Control
- https://educopia.org/research/etdplus

EDUCOPIA INSTITUTE

Photo by Simon Petrol on Unsplash

Learning Objectives

- Understand options for data management and data organization.
- Gain exposure to techniques and resources you may use to ensure your data will be readable and understandable in the future.
- Understand where to look for field-specific analysis methods, services, tools, and repositories.

Photo by Milada Vigerova on Unsplash

EDUCOPIA INSTITUTE

Key Takeaway

The decisions you make about how you organize and structure your data today will have implications for how you and others can access and make use (or sense!) of that data in the future.

Photo by Tim Gouw on Unsplash

What makes data hard to deal with?

- Data without data documentation (e.g., a data dictionary) is often impossible to understand.
- Without access to specific (often expensive) software, a data file may be unable to be viewed or used.
- IRB and funder requirements may impact the way you need to structure your data.
- As data usage increases, data often needs to be interoperable in order to enable sharing and reuse.

Photo by Markus Spiske on Unsplash

EDUCOPIA INSTITUTE

Questions to ask…repeatedly!

- What are the data organization standards for your field?
- What are the data export options in the software you are using?
- What forms of the data will be needed for future access?

Photo by Joe Shillington on Unsplash

EDUCOPIA INSTITUTE

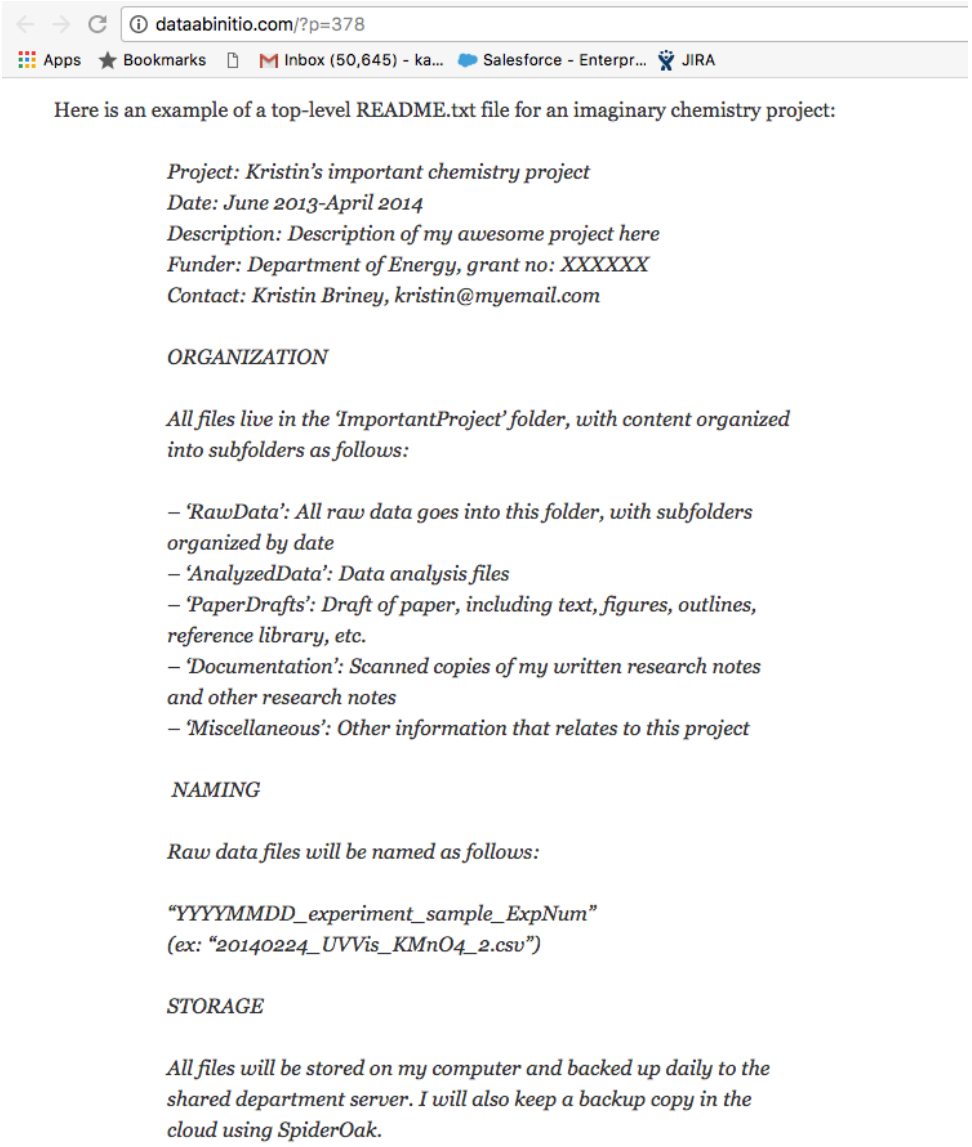
Structuring data well enables:

- Reproducible results
- Reuse in the future
- Sharing with others
- Gaining and retaining credibility
- Compliance with IRB/funder requirements

Providing Context for Your Data (cont)

Document:

- The data's purpose
- A list of the files in your data package
- See e.g., this ReadMe file example from Data Ab Initio



Providing Context for Your Data

Document:

- Data dictionary listing and describing all variables, e.g.:

(from DataOne and https://liberalarts.utexas.edu/redcap/_files/data_dictionary_example.jpg)

	A	B	C	D	E	
1	Variable / Field Name	Form Name	Section Header	Field Type	Field Label	Choices, Calculations, OR Slider Labels
2	participant_id	demographics		text	Participant ID	
3	enroll	demographics		text	Date subject signed consent	
4	fname	demographics		text	First Name	
5	lname	demographics		text	Last Name	
6	city	demographics		text	City	
7	state	demographics		text	State	
8	zip	demographics		text	Zipcode	
9	sex	demographics		dropdown	Gender	0, Female | 1, Male
10	given_birth	demographics		radio	Has the subject given birth before?	0, No | 1, Yes
11	num_children	demographics		text	How many times has the subject given birth?	
12	race	demographics		checkbox	Race	1, Caucasian | 2, African American | 3, Asian | 4, Other
13	race_other	demographics		text	Please describe:	
14	dob	demographics		text	Date of birth	
15	age	demographics		calc	Age	round(datediff([enroll],[dob],"y"),1)
16	height	demographics		text	Height (cm)	
17	weight	demographics		text	Weight (kilograms)	
18	bmi	demographics		calc	BMI	round([weight]*10000/([height]*[height]),1)
19	pcp	demographics		dropdown	Does patient have a primary care physician?	1, Yes | 2, No
20	upload	demographics		file	Upload record documents	

Data Organization Principles

- Use one variable per column.
- Make one observation per row.
- Use human-readable column names.
- Include one table per tab.
- Use a key to show relationships between tables

Movie Title	Director	Distributor	Running Time	Budget	Released
Peter Pan	Herbert Brenon	Paramount Pictures	105 minutes	40,030	Dec 29 1924
Girl Shy	Fred C. Newmeyer and Sam Taylor	Pathe Exchange	82 minutes	400,000	Apr 20 1924
Greed	Eric Von Stroheim	Metro-Goldwyn-Mayer	140 minutes	665,603	Dec 4 1924

Additional Principles

Do:

- Consider what your NULL values are and how they are represented
- Consider what contextual documentation is required
- Use standard data representations (e.g., (YYYYMMDD for dates)

Do Not:

- Use formatting to convey information
- Place comments in cells
- Use special characters in field names
- Use blank spaces or symbols in column names

Photo by Tim Gouw on Unsplash

Discipline-based data repositories:

- Social Sciences: ICPSR http://www.icpsr.umich.edu/icpsrweb/deposit/index.jsp
- Genomics: GenBank https://www.ncbi.nlm.nih.gov/genbank/
- Earth Sciences: NASA's Earthdata https://earthdata.nasa.gov/
- Archaeology: tDAR http://www.tdar.org/
- Oceanography: NODC http://www.nodc.noaa.gov/
- BioSciences: Dryad https://datadryad.org/

Data Organization

Structuring your data well enables you to:

- Reproduce results
- Reuse it in the future
- Share it with others
- Gain and retain credibility
- Comply with IRB/funder requirements

The decisions you make about how you organize and structure your data today will have implications for how you and others can access and make use (or sense!) of that data in the future.

Context and Data Documentation:

Include the following in a readme text file:

1. The data's purpose
2. A list of the files in your data package
3. Data dictionary listing and describing all variables

Data Organization Principles:

1. Use one variable per column
2. Make one observation per row
3. Use human-readable column name
4. Include one table per tab
5. Include an ID or key to indicate any relationship between tables

Whether your data is organized in lists, arrays, hash sets, dictionaries, queues, trees, heaps, or relational databases, it is important to be aware of disciplinary norms, as well as both institutional and funder requirements, that will make its deposit, storage, and long-term support more likely. Increasingly, the path for long-term support involves taking steps to make sure your data is deposited alongside data collected by others in your field or discipline.

Questions to consider for any data project:

1. What are the data organization standards for your field?
2. What are the data export options for your software?
3. What forms of the data will be needed for future access?

The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle.
https://www.dataone.org/best-practices

Do:

- Consider what your NULL values are and how they are represented
- Consider what data documentation is required
- Use standard data representations (e.g., (YYYYMMDD for dates)

***Do Not*:**

- Use formatting to convey information
- Place comments in cells
- Use special characters in field names
- Use blank spaces or symbols in column names

Discipline-based data repository examples:

-Social Sciences: ICPSR
-Genomics: GenBank
-Earth Sciences: NASA's Earthdata
-Archaeology: tDAR
-Oceanography: NODC
-BioSciences: Dryad

INSTITUTE of Museum and Library SERVICES

https://educopia.org/research/etdplus

EDUCOPIA INSTITUTE