Data Organization

Preservation and Curation of ETD Research Data and Complex Digital Objects

INSTITUTE of Museum and Library SERVICES

EDUCOPIA INSTITUTE

Photo by Tim Gouw on Unsplash

Data Organization

While conducting your research, you begin to amass a significant amount of information – survey responses, image files, interview transcripts, and geospatial data – that you plan to use in your thesis or dissertation, and that you may also want to reuse later in your career. What can you do to give yourself the best chance that your data will be findable and usable by both you and other researchers in the future? Considering how you organize, structure, and document your data is a good place to start. ■

Rationale and Motivations – Why

As you develop your research, you will have to consider how to manage and organize the data you're gathering. The type of research you are doing and the standards of your field(s) will affect your data organization and analysis practices. When you prepare and submit your final work as a thesis or dissertation, earlier decisions you have made about how the data is structured and organized will have implications for how readily others will be able to access and make use (or sense) of your data.

Each field has specific methods of analysis, as well as software tools that have been developed to help researchers accomplish that analysis. As you begin to gather, clean, and organize the data, consider not just how you will need to use the data today, but how to make sure it will be readable and understandable in the future.

Each discipline of study has data needs that are specific to the questions being asked. 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 institutional and funder requirements, which will make its deposit, preservation, 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.

Photo by Simson Petrol on Unsplash

The Basics – How to Do It

How researchers structure their data varies by disciplines and research questions. Still, there are general guidelines for structuring data that make it more likely to be usable in the future.

The following questions should be considered for any project that gathers data. These questions should be considered first at the planning stage, again as data is being gathered and stored, and once more prior to final deposit into a digital archive or repository.

1. **What are the data organization standards for your field?** For example, there are often standards for labeling data fields that will make your data machine-readable. There may also be specific variables and coding guidelines that you can use that will make your work interoperable with other datasets. Lastly, there may be accepted hierarchies and directory structures in your discipline that you can build upon.
2. **What are the data export options in the software you are using?** If using proprietary and/or highly specialized software to analyze large data sets, export the data in a format that is likely to be supported in the future, and that will be accessible from other software programs. This usually means choosing an open format that is not proprietary. Remember that you may not have access to the same software in the future, and not all software upgrades can read old file types.
3. **What forms of the data will be needed for future access?** Consider the various forms the data may take, and the scale of the data involved. You may need to preserve not only the underlying raw data, but also the resulting analyses you have created from it.

There is also a range of general principles that apply across many data types and forms that you can use to guide your work. These include the following:

Context and Data Documentation

In each folder or directory you produce, include a “readme” text file (i.e., a text-based file with the name “readme.txt”) detailing the following information:

- Abstract – describe why the data has been collected and for what purpose;
- Content – include a list of the files in your data package and a brief description of what each file is;
- Basic Data Dictionary – for each table (file) in the folder or data package, provide a list of the variables included in the file and a description of what each variable is.

Spreadsheet Structure

- Use one variable per column.
- Make one observation per row.
- Use human-readable column names.
- Include one table per tab.
- If you are using multiple related tables, use an ID or key to indicate how the tables are related.

Other Recommendations

Do:

- Consider what your NULL values are and how they are represented.
- Consider whether a more robust data dictionary is required (e.g. with more in-depth description of methods, instruments, models, etc. used to generate data).

Do Not:

- Use formatting to convey information.
- Place comments in cells.
- Use special characters in field names.

Table 1: Example Table Structure

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

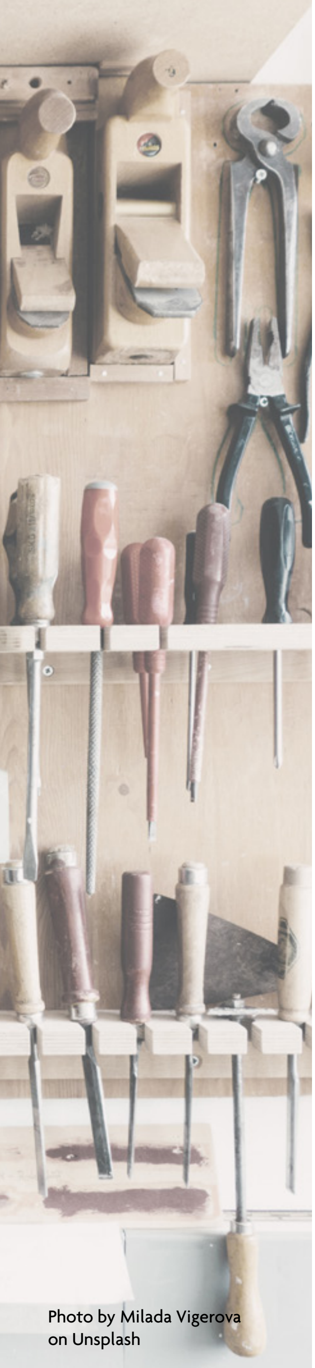

Photo by Milada Vigerova on Unsplash

Tools – What to Use

The "Basics" guidance above belies a tangle of disciplinary-specific guidelines for data curation, including structuring. Consult with your advisors, peers, and campus data specialists at the library to make sure you know the current state of guidance for your field. Some organizations provide useful links, including the Digital Curation Centre (DCC), which keeps an updated list of disciplinary metadata schemas.

- **Social Sciences:** The Data Documentation Initiative Alliance ("DDI") was created in 2003 as a collaborative effort to create metadata standards and semantic products for "describing social science data, data covering human activity, and other data based on observational methods." The organization is international in scope and includes academic associations, government organizations, and not-for-profit and for-profit organizations.
- **Archaeology:** This site offers a thorough discussion of the structure and types of data, as well as issues concerning long-term sustainability.
- **Geospatial Data:** Because of the way that many layers of data are used together to develop analyses, this report cautions that both the underlying data and resulting information graphics may need to be preserved. See Guy McGarva, Steve Morris, and Greg Janee, "Technology Watch Report: Preserving Geospatial Data," *Digital Preservation Coalition*, last updated May 2009.
- **Public Health:** This summary document provides an overview of the policies and standards encouraged within public health and epidemiology. See Wellcome Trust, "Enhancing Discoverability of Public Health and Epidemiology Research Data: Summary," *Public Health Research Data Forum*, last updated July 2014.

Resources

All links provided were last checked 10/3/2017, and the content we reference here was saved on that date in the Internet Archive's Wayback Machine. For links that no longer work, please visit the Wayback Machine and enter the url to surface the resource.

- The DataONE Best Practices database provides individuals with recommendations on how to effectively work with their data through all stages of the data lifecycle.
- Data Carpentry develops and teaches workshops on the fundamental data skills needed to conduct research, providing researchers with high-quality, domain-specific training covering the full lifecycle of data-driven research.
- The following introductory video series on data structures discusses how to store and organize data in a computer so that it can be efficiently used.
- For more information regarding how and why libraries increasingly accept ETD datasets, see W. Aaron Collie and Michael Witt, "A Practice and Value Proposal for Doctoral Dissertation Data Curation," *The International Journal of Digital Curation* 6, no. 2 (2011): 165-175.
- For a blog post that provides a detailed discussion of how to structure a dataset, including when to use particular tools and methods, see Diana Tkachenko, "Word Chains and Breadth-Wide Search," *The Garbage Collector* (blog), January 14, 2012.
- For a blog post that provides additional details on data structure, see Philippe Fournier-Viger, "Choosing data structures according to what you want to do," *The Data Mining & Research Blog* (blog), June 12, 2013.

Photo by Markus Spiske on Unsplash

Photo by Joe Shillington on Unsplash

Activities

1. Choose one spreadsheet you are using for a current data-gathering project.
   a. Using the section above titled “Structure,” check to see if your file meets those requirements.
   b. Create a data dictionary for the spreadsheet that describes the meaning of each column header.

INSTITUTE of Museum and Library SERVICES

EDUCOPIA INSTITUTE