

Data Management Workshop 2: Data Documentation

Elizabeth Wickes, Data Curation Specialist (wickes1@illinois.edu)

Ayla Stein, Metadata Librarian (astein@illinois.edu)

Peg Burnette, Biomedical Sciences Librarian (phburn@illinois.edu)

Susan Braxton, Prairie Research Institute Librarian (braxton@illinois.edu)

About this workshop

Writing project, code, and data documentation doesn't need to be the worst part of your day. This hands on workshop will give you experience using various types of documentation, discuss strategies for writing documentation, and get you started writing a template for your projects. Bring a dataset you'd like to work with but examples will be provided. Materials will be available at:

Upcoming workshops

Preparing for Data Sharing (Data Management Workshop 3)

September 27, 10-11am, Main Library 314

Registration link: <http://illinois.edu/calendar/detail/4068?eventId=33240140>

Making research data public is becoming a reality for many disciplines, but for many researchers and disciplines there is a complicated set of issues to consider before publication or sharing data. This workshop will cover the basic steps of research data publication, from considerations to depositing. Participants will work through guidance to help them make decisions about when and how to publish or share data.

Data Rescue! (Data Management Workshop 1)

October 4, 12-1pm, Grainger Library 335

Registration link: <http://illinois.edu/calendar/detail/4068?eventId=33240142>

Have a mess of data floating around your computer or lab? Data management is an essential task for students and faculty but hard to get started. Go no further! In this workshop you will identify, group, and plan on how to reorganize your current and future data. We will discuss strategies for organization, folder structure, and create an organizational plan.

Need more help after this workshop?

The Research Data Service (<http://researchdataservice.illinois.edu/>) team is happy to provide more information about all data management topics and offers personal data consultations. They can be reached at researchdata@library.illinois.edu or on Twitter @ILresearchdata. This workshop and many others can be run for your research group or lab by request. All Research Data Service consultations are free and confidential to all members of the Urbana-Champaign campus community.

Creating Data Documentation: Activity 1 – Using Documentation

Step 1: Go to these dataset pages:

1. Han, Xueying; Appelbaum, Richard; Stocking, Galen; Gebbie, Matthew. International STEM Graduate Student in the United States Survey 2015. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-08-10. **<http://doi.org/10.3886/E43668V1>**
2. Meili, Stephen. Do Human Rights Treaties Help Asylum-Seekers: Findings from the U.K.. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2015-05-21. **<http://doi.org/10.3886/E17507V2>**

You'll need to make an account with them to download, but we can provide you with digital copies if you need them.

Step 2: Take no more than 7-10 minutes for this activity – the process is what's important, not the answers! Each of these problems can be solved, but there are some curveballs. You may need to dig into other resources to answer them.

- Review the variables measured in the data and attempt to answer a few questions about it. You may pick something that is of interest to you, or attempt to find the participant's demographic details. For example, find the total sample size and the gender breakdown for the respondents who participated in each study.
 - Han, Applebaum, Stocking, & Galen (2015):

 - Meili (2014):
- What uncertainties, questions, or confusion points did you encounter in determining your answers?

- What did you find helpful, convenient, or crucial in determining your answers?

- What documentation could you have taken away and still known enough to answer this question?

Activity 2: Sketching out what your documentation could look like

Step 1: Try to think of a specific dataset you are working with. You may also answer these questions for the general type of data that you work with. Write the name in this space.

Step 2: Identify the people who may need to use your data or project materials. Check all that apply. You may have multiple checkmarks per audience.

Currently	Within the next 2 years	2-5 years from now	5+ years	Audience
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Just me
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	My project advisor or PI
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Others in my project team
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	My lab or department members
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Researchers in my field
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Researchers outside of my field
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Publication reviewers
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Federal or government agency
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	_____

Step 3: Use the tables below to identify the things that make up your project.

Are there any tools, devices, or other physical materials or objects (including paper surveys, paper coding sheets, audio tape recordings, etc.) this project depends on? Use the space below each question to write down information about them. Refer to your Workshop 1: Data Rescue materials if you completed them.

Used to gather the data:

Used to process the data:

Necessary to understand the data:

--	--	--

Think about all the components within your workflow, including items required to access, transcribe, process, and analyze your data. Write them down here under the prompts. Change the questions or labels as necessary for your project.

Data files

What is it?	File name?	Where is it?	Who made it?

Samples or other physical objects, including paper surveys:

What is it?	ID?	Where is it?	Who made it?

Code or scripts:

What is it?	File name?	Where is it?	Who made it?

Other/Extra space:

What is it?	Name or ID?	Where is it?	Who made it?

Step 4: Identify the relevant sections for your data or project. Check relevant items. When thinking of how important each section is, consider these sections in the context of the timelines identified in step 2.

- Administrative and personnel details:
 - Authors, principle investigators, contributors, etc., with associated including institutions, contact information, and other identifiers when available
 - Description of project team
 - Associated papers, code, talks, datasets, etc.
 - Funders and grant numbers
 - Data licensing information
 - Suggested citation
 - Who to contact if there are any questions
- Project information:
 - Brief description of the dataset and/or abstract, including relevant collection and processing dates
 - Collection methods, including dates of collection, data processing, etc.
 - Names, model numbers, and calibration information for any instruments used during data collection
 - Description of scripts (e.g. R, Python, MATLAB, etc.) and their purpose
 - Data processing workflow and stages
 - Data cleaning process
 - De-identification or other data scrubbing steps that occurred
- Data file information:
 - List of files to be included, grouped in meaningful units
 - Number of rows, fields, columns, etc.
 - Description of folder contents and/or of large groups of similar files
 - Explanation of formats and required software to read them
 - Languages represented within your data
 - Description of the values, units, etc. for each column or field (codebook)
 - Description of columns and fields in the data files (data dictionary)
 -
 - Other domain specific descriptive information
- Other:
 -
 -
 -
 -
 -

Homework!

Continue working through these elements and begin sketching out what you know, what you need to look up, and anything that someone else needs to write.

Step 5: Review some examples and other documentation. Skimming is fine!

- Cornell's "Guide to writing "readme" style metadata"
 - <http://data.research.cornell.edu/content/readme>
- ICPSR's Data Preparation Guide: Important Metadata Elements (Social science)
 - <https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html>
- Find a data repository with data in your research area. Review some of their guides or a few popular deposits.
 - Registry of Research Data Repositories - <http://www.re3data.org/>

Step 6: Start writing.

- Start with outlining the sections you selected in Step 4.
- Transcribe the information you wrote down in Step 3 into the relevant sections.
- Consider the timeline and audience that you selected in Step 2. Make new sections or add details necessary for shorter and longer-term use as necessary.
- Start filling in the details!

Other Important Considerations

- You can write your documentation with the presumption that future users (including yourself in the future) have some domain specific knowledge.
- Visualizations and other infographics or diagrams are fine to use! Particularly for workflow mapping, file structure descriptions, etc.
- Create a template documentation file for your common types of projects, and creating an internal style guide if you are expecting to have students or other researchers prepare this documentation.
- Start the documentation process early in the project while you have all project members around to gather information from, and while the information is fresh in your mind.
- Add links to articles, presentations, etc. when they contain essential information. Keep local copies of each in case you lose access.
- Create documentation boilerplate language for methods and tools that are shared between several projects.
- Have someone review your documentation. Having someone knowledgeable about the project review content and someone who doesn't for clarity can often make the strongest documentation.

Additional Resources

Readme Files

- Cornell's "Guide to writing "readme" style metadata"
 - <http://data.research.cornell.edu/content/readme>
- ICPSR's Data Preparation Guide: Important Metadata Elements (Social science)
 - <https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html>
 - Pay special attention to the "Important metadata elements section".

Data Dictionaries & Codebooks

- Creating a Codebook, the Center for Criminal Justice Research:
 - <http://ccjr.csusb.edu/docs/researchmanualdocs/creatingacodebook.pdf>
- Codebook cookbook: How to enter and document your data
 - <http://www.medicine.mcgill.ca/epidemiology/joseph/pbelisle/CodebookCookbook.html>
- Data Ab Initio - Data Dictionaries
 - <http://dataabinitio.com/?p=454>