Guide to Creating Documentation for Your Dataset in Deep Blue Data

Research Data Services – researchdataservices@umich.edu

Why documentation is important

To fully understand, trust, and make use of a dataset created by someone else, researchers often need more than the data files: they need a detailed understanding of the context in which the dataset was created; the methods used to collect, process, and analyze the data; and what the values and terms in the data signify. Best practices in research data curation point to including documentation within a dataset itself as a dependable way to convey comprehensive information about the data, and the research that produced it, to interested members of academic communities and the general public.

Some of this information may be provided in journal articles, monographs, or other texts reporting research findings. However, because they are typically written to describe the methods and results, research publications often leave out key details about the data itself, including how data files are structured; definitions of headings, variables, and other terms used in the data; and how the individual components of the dataset fit together. Even when this information is provided in a research publication, the text may not be easily accessible to those seeking to understand and make use of the data.

What we mean by documentation and metadata

Some data curation experts and practitioners employ the terms documentation and metadata interchangeably. Deep Blue Data chooses to differentiate between the two as, in the context of the repository, they serve different purposes.

Metadata here refer to the high-level information that researchers provide when depositing data in Deep Blue Data and include the Title, Creator, Description, Methods, Keywords, and other descriptive information. Metadata help people who may be interested in a dataset to find it and understand it well enough to decide whether or not they would like to download it from Deep Blue Data.

The information included in documentation is more expansive than what is provided in the metadata. While metadata helps users to discover and identify datasets of interest, documentation provides a richer understanding and establishes trust in the data. Ideally, the documentation would be comprehensive enough to enable others to reuse the data for new projects or to reproduce the research to verify its findings. Documentation can come in many forms depending on the nature of the data and the research; examples include a codebook, a field notebook, or a README file.

What we recommend including in documentation

To facilitate usability of the datasets on Deep Blue Data, Research Data Services requests that researchers include documentation to accompany the data either in separate files or embedded within the data files themselves. Ideally, documentation would be produced throughout the research process, making it easy to compile and combine with the data before deposit, but we recognize that this does not always happen. Research Data Services offers educational programming and consultation on developing documentation to accompany a dataset deposited into Deep Blue Data or elsewhere. Please contact us if you would like to take advantage of our services.

While the practice of creating documentation for data is broadly supported (see bibliography below), the recommended format and content varies between fields, datasets, and repositories. We would be happy to work with you to identify best practices for documenting data in your field of study and to consider how these practices might be applied to your dataset.

Though we acknowledge practices and needs for describing research data will be determined case by case, the Research Data Services team feels that the following areas should be covered at minimum in documentation:

- Research Overview: A summary of the subject and purpose of the research, as well as who conducted the research, where and when it took place, and the research funding sources.
- Methods: An account of how the research was conducted, with a focus on how data were collected, processed, and analyzed. Sampling procedures, instruments, and software used should be described to convey how data were produced and transformed.
- File Inventory: A list of each of the files included in the dataset, including a brief statement of what each file contains and its purpose. An explanation of the organization of the files, including relationships between files, should be considered as well.
- Definition of Terms and Variables: A glossary that specifies clearly and without jargon the meaning of ambiguous terms, obscure procedures, variables, and/or units that appear throughout the dataset. This may take the form of a list of definitions, a data dictionary in a spreadsheet, or comment lines distributed through a program.
- Use and Access: (if needed) Instructions on how to open, run or make use of the data files.
- Suggest a Citation for the Data Set: (if needed) Adding a suggested citation for the data set will encourage people to give you attribution for your work.

Further reading

Research Data Services has created a bibliography of resources to assist researchers in understanding and creating documentation. The resources originate from academic institutions, researchers, repositories, and other organizations. Some resources below approach the data documentation topic generally, while others apply to specific kinds of data or research.

General Resources

Cornell University. n.d. "Guide to writing 'readme' style metadata." Research Data Management Service Group. Accessed Jan. 26, 2018. https://data.research.cornell.edu/content/readme#bestpractices.

Cornell University's Research Data Management Service Group has a page sharing detailed instructions that cover the above recommended topics and more. The group also has a plain text document template that can serve as a useful starting point and be modified as needed.

Dryad. 2018. "Best practices for creating reusable Dryad data packages." Last revised on Jan. 24, 2018. http://datadryad.org/pages/reusabilityBestPractices.

Dryad Digital Repository, a nonprofit research data repository that supports the scientific and medical disciplines, recommends the creation of README documents at the file or "data package" levels. This page outlines content to include in these documents and provides links to examples in the repository.

Open Science Framework. 2017. "How to Make a Data Dictionary." OSF Guides. Updated on April 4, 2017. http://help.osf.io/m/bestpractices/l/618767-how-to-make-a-data-dictionary.

This Open Science Framework Guide provides column-by-column instructions on creating a data dictionary — a common way to document the use and meaning of variables and headers in tabular data files.

Strasser, Carly, Robert Cook, William Michener, Amber Budden. n.d. "Primer on Data Management: What you always wanted to know." Data Observation Network for Earth (DataONE). Accessed Jan. 29, 2018. https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf.

DataONE — an organization facilitating discovery of Earth and environmental science research data and educating researchers on best practices — gives an introduction to the steps of the data management life cycle. Sections 5.2 and 5.4 are especially useful for information about documenting data and the research process. DataONE also has an extensive list of Best Practices,

many of which relate to documentation, available here (embed link: https://www.dataone.org/all-best-practices).

UK Data Archive. n.d. "Document your data." Manage data. Accessed Jan. 26, 2018. https://www.ukdataservice.ac.uk/manage-data/document.

The UK Data Archive, a repository for social and economics data, provides recommendations on preparing documentation at the study and data levels, with specific guidance for quantitative and qualitative data.

Resources Specific to Disciplines or Data Types

Archaeology Data Service and Digital Antiquity. n.d. "Project Documentation." Guides to Good Practice. Accessed Feb. 19, 2017. http://guides.archaeologydataservice.ac.uk/g2gp/CreateData_1-1.

This webpage, part of a resource created by these archaeological data authorities in the United Kingdom and United States, include documentation guidance tailored to archaeological research and datasts. It addresses how to document stratigraphic units, codes used, and drawings, among other parts of an archaeological data workflow.

ICPSR (Inter-university Consortium for Political and Social Research). 2012. *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle* (5th ed.). Ann Arbor, Michigan: ICPSR. https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.

A leader in the data curation field, ICPSR publishes a guide (in PDF format) on preparing data for deposit that focuses on research data from the social sciences. Most notably, "Phase 3: Data Collection and File Creation" (p. 21-32) includes discussion of creating documentation for quantitative and qualitative data.

Jackson, Mike. n.d. "Writing readable source code." Software Sustainability Institute. Accessed Jan. 26, 2018. https://software.ac.uk/resources/guides/writing-readable-source-code?_ga=2.180309094.396671004.1516810225-231542810.1516810225#node-131.

For the Software Sustainability Institute, Mike Jackson writes about best practices for making source code readable by humans; he covers comments, formatting, available tools for creating documentation and maintaining code conventions, and more.

U.S. National Library of Medicine. 2016. “Common Data Element (CDE) Resource Portal.” Updated March 29, 2016. https://www.nlm.nih.gov/cde/.

The National Institutes of Health encourages the use of Common Data Elements (CDEs), which are common in medical fields to facilitate data quality, data sharing, and comparison across studies. This portal links to a suite of resources for determining CDEs for a field or subfield.

Wilson, Greg, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, et al. 2014. “Best Practices for Scientific Computing.” *PLOS Biology* 12, no. 1: e1001745. https://doi.org/10.1371/journal.pbio.1001745.

Taking a broad approach to improving programming practice in the sciences, this article includes suggestions on code structure, documentation of the environment, the use of version control, and writing comments to enhance research reproducibility.

Have Questions? Need Help?

Please contact Research Data Services - researchdataservices@umich.edu