Data documentation and metadata

Andrew Johnson
Research Data Librarian
andrew.m.johnson@colorado.edu

Earth Lab, July 26, 2016

Outline

- What are data documentation and metadata and why are they important?
- Good practices for creating data documentation and metadata
- Examples from the wild
- Q&A

What are data documentation and metadata?

Data documentation

Describes the who, what, where, when, and how surrounding data creation/collection so that others outside of the project can understand and reuse data

Metadata

Describes the who, what, where, when, and how surrounding data creation/collection so that others outside of the project can *discover*, understand, and reuse data.

Typically machine-readable, structured, and standards-based.

Why are data documentation and metadata important?

“Metadata is a love note to the future”

- But, who will read it?
  - Your future self?
  - Your colleagues?
  - The broader research community?
  - The general public?
- What will they need to know?
  - To find and access your data?
  - To understand your data and how it was created/collected?
  - To reuse your data?
- Helpful to start at the end
  - Where will your data eventually live?
  - Does that location provide guidelines/examples?

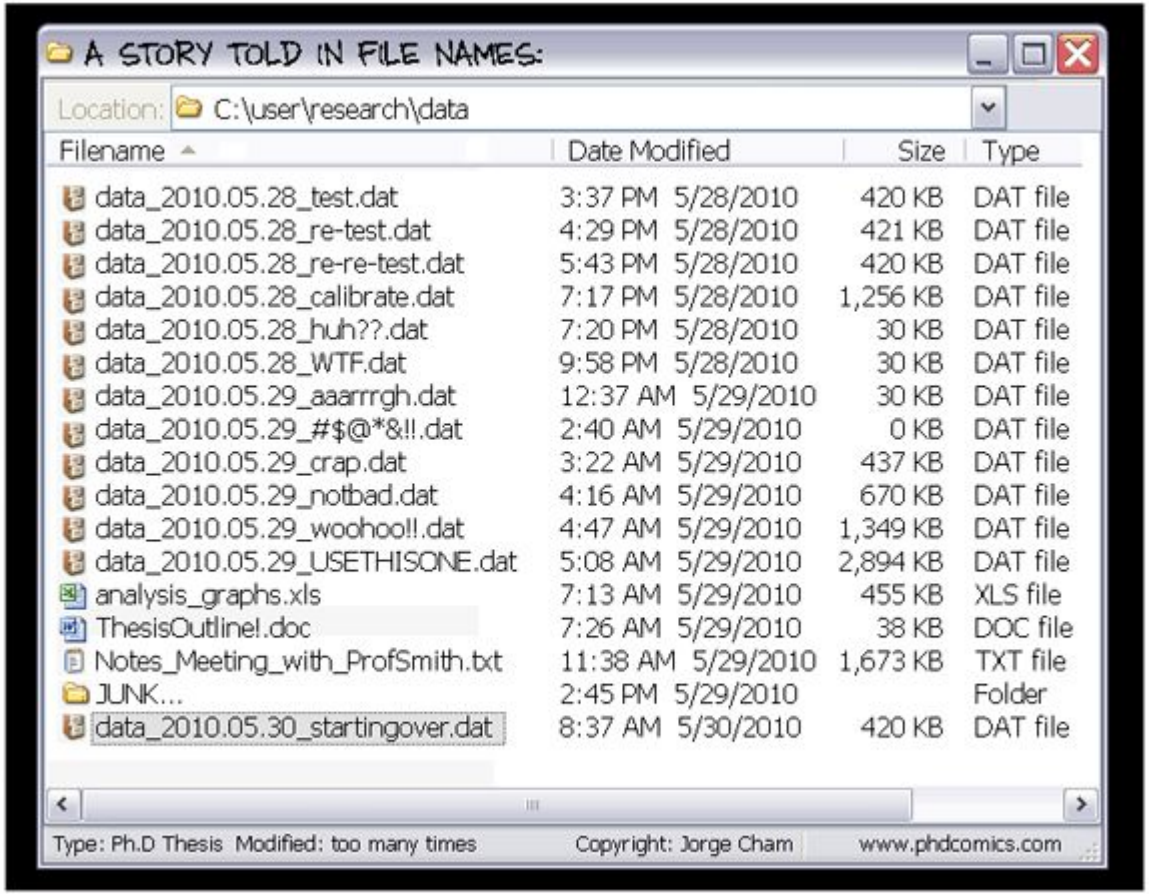

Image source: http://www.phdcomics.com/comics/archive.php/tellafriend.php?comicid=1323

Good practices for creating data documentation and metadata

Describe project

- In a project-level “readme” file (plain text) include:
  - Basic project info (title, institution(s) involved, source(s) of funding)
  - Project personnel (principal investigator, researchers, technicians, others) and contact info
  - Location and description of study site or sites
  - Range of dates for the project
  - Rationale for the project
  - Description of project methods
  - Licenses or restrictions placed on data
  - Related resources
  - Recommended citation: Author(s), Year, Title, Repository or Archive, Version, Identifier

Describe how data are organized

- Describe where and how to access all data
- Use a logical structure to organize data files/directories
- Reflect this structure in file/directory names
  - May include project name, location, investigator name, date, data type, version, etc.
  - Use lower-case in general
  - Avoid spaces and special characters
- Document this structure in the project-level “readme” file

Describe data

- In a separate “readme” for each data file or data set:
  - Define parameters:
    - Use standard names across files, data sets, projects
    - Include parameter name, how it was measured (including units), and abbreviation used (if applicable)
    - Do not abbreviate units
  - Describe formats for dates, times, geographic coordinates, etc. (e.g., ISO 8601 for dates/times: https://www.w3.org/TR/NOTE-datetime)
  - Define any coded values
  - Define missing values (e.g., -9999) and notes about why data is missing
  - Describe any quality or other issues with data
  - Clearly identify any existing data sets used and steps taken to integrate or derive data
  - Provide versioning information

If possible: Use standardized vocabularies

- Integrated Taxonomic Information System (taxonomic information): http://www.itis.gov
- NASA Thesaurus (engineering, physics, space sciences, earth sciences): http://www.sti.nasa.gov/sti-tools
- GCMD Keywords (earth and climate sciences, instruments, sensors, data centers, etc.): http://gcmd.nasa.gov/learn/keywords.html
- USGS Biocomplexity Thesaurus (agriculture, forest, fisheries, etc.): https://www2.usgs.gov/core_science_systems/csas/biocomplex_thesaurus

If needed: Create standards-based metadata

- Examples:
  - DataCite (general): https://schema.datacite.org/
  - Ecological Metadata Language (ecology): https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
  - Data Documentation Initiative (social sciences): http://www.ddialliance.org/
  - ISO 19115 (geospatial): http://www.fgdc.gov/metadata/geospatial-metadata-standards
  - Darwin Core (biodiversity): http://rs.tdwg.org/dwc/
- Useful to use tools to create:
  - ISO geospatial metadata editors: https://www.fgdc.gov/iso-metadata-editors-registry/editors
  - Morpho (EML metadata editor): https://knb.ecoinformatics.org/#tools/morpho
  - DDI Metadata Editor: http://www.ihsn.org/home/software/ddi-metadata-editor

Quality control

- Have a “naive” user inspect documentation and/or analyze data
- Does the documentation accurately describe the data?
- Are there errors or is anything missing from the documentation?
- Can a task (e.g., data analysis) be successfully completed using only the data and metadata?

Examples

- Bond-Lamberty, B.P. and A.M. Thomson. 2014. A Global Database of Soil Respiration Data, Version 3.0. Oak Ridge, Tennessee USA. Oak Ridge National Laboratory Distributed Active Archive Center. doi: http://dx.doi.org/10.3334/ORNLDAAC/1235
- Fetterer, F., K. Knowles, W. Meier, and M. Savoie. 2016, updated daily. Sea Ice Index, Version 2. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. doi: http://dx.doi.org/10.7265/N5736NV7
- DataCite example metadata record: https://schema.labs.datacite.org/meta/kernel-4.0/example/datacite-example-dataset-v4.0.xml

Questions?

Acknowledgments

This work was adapted in part from the following guides:

- Cornell University Research Data Management Service Group. *Guide to Writing "readme" Style Metadata*. http://data.research.cornell.edu/content/readme
- DataONE. *Best Practices*. https://www.dataone.org/best-practices
- University of Minnesota Libraries. *Data Documentation and Metadata*. https://www.lib.umn.edu/datamanagement/metadata