Take note(s): Good practices for data documentation

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

Earth Lab, February 22, 2017

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

A.k.a. 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 is data documentation 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?
  - http://www.nature.com/sdata/policies/repositories

Good practices for creating data documentation

In a README: High level info

- Title
- Source(s) of funding
- Personnel and contact info (including institutional affiliation)
- Geographic location(s)
- Date(s)
- Licenses or restrictions placed on data
- Related resources (publications, other data sets, software, etc.)
- Version, other locations
- Recommended citation:
    - Author(s), Year, Title, Repository or Archive, Version, Identifier

In a README: Nitty gritty

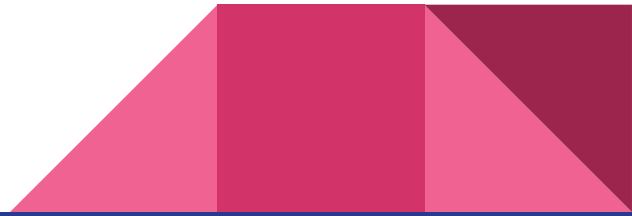
- File list and relationships
- Methodological information
- Parameters:
  - Use standard names across files, data sets, projects
  - Include parameter name, how it was measured (including units), and abbreviation used (if applicable)
  - Do not abbreviate units
- Formats for dates, times, geographic coordinates, etc. (e.g., ISO 8601 for dates/times: https://www.w3.org/TR/NOTE-datetime)
- Coded values
- Missing values (e.g., -9999) and explanations
- Any quality or other issues with data

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/biocomplexity_thesaurus/

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

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