Guide for writing “Readme” files

A readme file provides information about a data file and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data. Standards-based metadata is generally preferable, but where no appropriate standard exists, for internal use, writing “readme” style metadata is an appropriate strategy.

Want a template? Download one and adapt it for your own data: cornell.box.com/v/ReadmeTemplate

Best practices

Create readme files for logical “clusters” of data. In many cases it will be appropriate to create one document for a dataset that has multiple, related, similarly formatted files, or files that are logically grouped together for use (e.g. a collection of Matlab scripts). Sometimes it may make sense to create a readme for a single data file.

Name the readme so that it is easily associated with the data file(s) it describes.

Write your readme document as a plain text file, avoiding proprietary formats such as MS Word whenever possible. Format the readme document so it is easy to understand (e.g. separate important pieces of information with blank lines, rather than having all the information in one long paragraph).

Format multiple readme files identically. Present the information in the same order, using the same terminology.

Use standardized date formats. Suggested format: W3C/ISO 8601 date standard, which specifies the international standard notation of YYYY-MM-DD or YYYY-MM-DDThh:mm:ss.

Follow the scientific conventions for your discipline for taxonomic, geospatial and geologic names and keywords. Whenever possible, use terms from standardized taxonomies and vocabularies, a few of which are listed below.

Source	Content	URL
Getty Research Institute Vocabularies	geographic names, art & architecture, cultural objects, artist names	http://www.getty.edu/research/tools/vocabularies/
Integrated Taxonomic Information System	taxonomic information on plants, animals, fungi, microbes	http://www.itis.gov/
NASA Thesauri	engineering, physics, astronomy, astrophysics, planetary science, Earth sciences, biological sciences	https://www.sti.nasa.gov/nasa-thesaurus/
GCMD Keywords	Earth & climate sciences, instruments, sensors, services, data centers, etc.	https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords
The Gene Ontology Vocabulary	gene product characteristics, gene product annotation	http://amigo.geneontology.org/amigo/dd_browse
USGS Thesauri	agriculture, forest, fisheries, Earth sciences, life sciences, engineering, planetary sciences, social sciences etc.	https://www1.usgs.gov/csas/biocomplexity_thesaurus/index.html
IUPAC Gold Book	compendium of chemical terminology from the International Union of Pure and Applied Chemistry (IUPAC)	https://goldbook.iupac.org

Recommended content

Recommended minimum content for data re-use is in **bold**.

General information

1. **Provide a title for the dataset**
2. **Name/institution/address/email information for**
    - **Principal investigator (or person responsible for collecting the data)**
    - Associate or co-investigators
    - Contact person for questions
3. **Date of data collection (can be a single date, or a range)**
4. **Information about geographic location of data collection**
5. Keywords used to describe the data topic
6. Language information
7. Information about funding sources that supported the collection of the data

Data and file overview

1. **For each filename, a short description of what data it contains**
2. Format of the file if not obvious from the file name
3. If the data set includes multiple files that relate to one another, the relationship between the files or a description of the file structure that holds them (possible terminology might include “dataset” or “study” or “data package”)
4. **Date that the file was created**
5. Date(s) that the file(s) was updated (versioned) and the nature of the update(s), if applicable
6. Information about related data collected but that is not in the described dataset

Sharing and access information

1. **Licenses or restrictions placed on the data**
2. Links to publications that cite or use the data
3. Links to other publicly accessible locations of the data (see best practices for sharing data for more information about identifying repositories)
4. Recommended citation for the data (see best practices for data citation)

Methodological information

1. **Description of methods for data collection or generation** (include links or references to publications or other documentation containing experimental design or protocols used)
2. **Description of methods used for data processing (describe how the data were generated from the raw or collected data)**
3. Any software or instrument-specific information needed to understand or interpret the data, including software and hardware version numbers
4. Standards and calibration information, if appropriate
5. Describe any quality-assurance procedures performed on the data
6. Definitions of codes or symbols used to note or characterize low quality/questionable/outliers that people should be aware of
7. People involved with sample collection, processing, analysis and/or submission

Data-specific information

Repeat this section as needed for each dataset (or file, as appropriate)

1. Count of number of variables, and number of cases or rows
2. **Variable list, including full names and definitions (spell out abbreviated words) of column headings for tabular data**
3. **Units of measurement**
4. **Definitions for codes or symbols used to record missing data**
5. Specialized formats or other abbreviations used

Want a template? Download one and adapt it for your own data: cornell.box.com/v/ReadmeTemplate

References

The preceding guidelines have been adapted from several sources, including:

Best practices for creating reusable data publications. Dryad. 2019. https://datadryad.org/stash/best_practices

Introduction to Ecological Metadata Language (EML). The Knowledge Network for Biocomplexity. 2012. https://web.archive.org/web/20120424124714/http://knb.ecoinformatics.org/eml_metadata_guide.html