

# Smithsonian Data Management Best Practices

## Describing Your Data: Data Dictionaries

A "data dictionary" or a "readme" file includes crucial information about your data that ensures it can be correctly interpreted and re-used by yourself, possible collaborators, and other researchers in the future. Depending on the nature of your datasets<sup>iii</sup>, it may include collection methods or any processing/calculations that were applied to the dataset as a whole or to specific data elements.

*"The increased use of data processing and electronic data interchange heavily relies on accurate, reliable, controllable, and verifiable data recorded in databases. One of the prerequisites for a correct and proper use and interpretation of data is that both users and owners of data have a common understanding of the meaning and descriptive characteristics (e.g., representation) of that data. To guarantee this shared view, a number of basic attributes has to be defined."*

*-International Standards Organization (ISO) Information Technology Parts 1-6 (2nd Edition), 2004.*

If the data you are describing is primarily tabular, the description could be in a tabular form as well. In most cases, you should create your dictionary as a **plain text file** with an introduction giving basic information about the dataset, followed by detailed definitions for each element in the dataset.

- Create one descriptive file (dictionary) for each dataset.
- Name the dataset, data dictionary, and any other supporting files similarly.
- Start with basic information about the data, the same information found in a citation record.<sup>iv</sup>
- See the best practices for *Working with Tabular Data* and *File Naming* for more details

Follow the conventions of your discipline when choosing standardized terms or when structuring your data, e.g., use USGS Thesauri terms for Earth science data, or Darwin Core for Natural History collections.

You should also provide sufficient metadata to cite your dataset (see Describing Your Project: Citation Metadata). This information may be included in the data dictionary, or be stored separately in another file, or as a metadata record in a repository.

A basic overall definition of the data should be at the beginning of your data dictionary.

**Basic dataset introduction** *must* include:

- Who collected or aggregated that data, or in the case of many contributors, who is the principal investigator or contact.
- When the data was collected.
- What the data elements are measuring or describing.
- Why the data was collected.
- Methodologies used or assumptions made while collecting the data.

Additional definition *should* include when relevant:

- Description of any transformations or calculations applied to the raw data (if the data being described is not the raw data) or to specific data elements, including references to any scripts used.
- Version<sup>v</sup>.
- Any validation or quality control process that has been applied.

After the overall definition, describe each component or element of your data. If your data is tabular, describe each column (field) and what it should contain. If your data includes images, describe how they are organized, and where detailed metadata can be found.

Data element description may include:

- Element name as found in the dataset, i.e., the data label, column header, or filename.
- A full “human readable” name of the element if the dataset uses codes or abbreviations.
- A definition of the data element.
- Any units of measure and precision (if applicable), e.g., “measured in meters, rounded up to the nearest .01 meter”.
- The format of the data element (if applicable), such as integer, text, date-time, etc.
- All valid/allowed values.
- Any codes, symbols or abbreviations used in the values themselves.
- If the element is required or not.
- The source of the controlled vocabulary or thesaurus used (if applicable).
- The source of the data element, e.g., sensor, observation, etc.
- If the element is “null”(for a non-required element) the convention for how that is represented, e.g., “unk”, “-999999”

## Example

File 1: Amendment seed packets and fungi\_all.txt

This datafile includes the numbers of protocorms recovered from seedpackets exposed to amendment with different organic amendments, compared to no amendment. Data were collected 2010-04-02 and 2010-04-08 with results published in the paper “title of paper.” Missing data are indicated by a “.”. Data were collected by M----- and R-----. Questions should be directed to M-----.

Column headings:

Species: The orchid species of seeds added to the plot in seedpacket. Goodyera=Goodyera pubescens; Liparis=Liparis liliifolia; Tipularia=Tipularia discolor

Site: Designated numerically 1-6. All sites are forest stands at the Smithsonian Environmental Research Center, Edgewater, Maryland, USA. Sites 1-3 are old stands and 4-6 are young stands (see Siteage, below).

Subplot: Designates the subplot location within each site. Thirty-six subplots were arranged in a square with columns labeled A-F and rows labeled 1-6.

Siteage: Old=120-150year old forest. Young=50-70year old forest.

Treatment: The amendment added to a subplot (Leaves=tulip poplar leaf litter; Wood=chipped fresh tulip poplar wood). Subplots with no amendment added are designated Control.

Inoculated?: Designates whether mycorrhizal host fungi were inoculated into the subplot.

fungusyn: Indicates whether appropriate host fungi were detected (1) or not (0) using PCR amplification of the soil in the subplot.

fungusInt: A semi-quantitative measure of the abundance of appropriate host fungi. The intensity of fluorescence by a post-PCR gel band 0=no band visible to 3=intensely bright fluorescence.

fung2YN: For Tipularia discolor, indicates whether an appropriate host fungus was detected (1) or not (0) using PCR amplification of the soil in the subplot using a second primer set (TipC2F/TipR) that detects an appropriate host fungus not detected by the first primer set (TipC1F/TipR).

## References

2006. Northwest Environmental Data Network. Best Practices for Data Dictionary Definitions and Usage.

[http://www.pnamp.org/sites/default/files/best\\_practices\\_for\\_data\\_dictionary\\_definitions\\_and\\_usage\\_version\\_1.1\\_2006-11-14.pdf](http://www.pnamp.org/sites/default/files/best_practices_for_data_dictionary_definitions_and_usage_version_1.1_2006-11-14.pdf)

Retrieved 2017. Open Science Framework. How to make a data dictionary. <http://help.osf.io/m/bestpractices/l/618767-how-to-make-a-data-dictionary>

2017. USGS. Data Management: Data Dictionaries and Thesauri.

<https://www2.usgs.gov/datamanagement/describe/dictionaries.php>

---

<sup>i</sup> Data dictionary is a file that describes each element of your dataset. If your dataset includes tabular data, R code, and images, the data dictionary would include a list of the fields in the table and what they mean, including units and precision; a brief overview of the purpose of the code (if not already contained in comments); and information about the images and how they relate to the dataset (more detailed metadata for the images should be embedded). For a brief guide see <http://datadryad.org/pages/readme>

<sup>ii</sup> Readme – another term for data dictionary, however Readme files usually are narrative rather than contain detailed metadata.

<sup>iii</sup> A dataset is a logical grouping of files or data points that support, or are used as evidence in, a specific research enquiry. Datasets may include one tabular file, a set of 10,000 images, R code, or combinations of multiple types of files.

<sup>iv</sup> Citation metadata or metadata record contains the overall description of your dataset (or project), sufficient to find and properly cite your dataset (or project).

Essential components of a dataset citation include: Creator/owner (with ORCID); Title; Date made public; DOI (or other persistent id); and, preferably, an abstract which provides a brief narrative describing the data (not identical to the paper abstract, if data is supporting a paper.)

<sup>v</sup> Versioning is the process of saving, and then assigning unique names or numbers, to a file whenever significant changes are made. Versioning files enables tracking changes over time, as well as supporting reverting to an older version of a file. Several software tools and file systems support automatic versioning including: DropBox, Git, ArcGIS, and most wikis.