Smithsonian Data Management Best Practices

Working with Tabular Data

Putting data into simple tables is one of the most common ways to store and then work with data. Below are some basic principles for organizing data into tables so that both humans and machines can use that data.

If your project or group already follows a convention for putting data in tabular form, always follow that convention. However, if that convention is significantly different from one of the guidelines below, *consider retaining a copy of the data in its original form as well as a normalized version* that conforms to SI guidelines.

General guidelines

- If possible, store tabular data in a non-proprietary file format such as comma-delimited .csv or tab-delimited .txt files.
- Do not rely on special formatting such as cell colors, text bolding, or other visual cues to provide meaning.
- Do not include figures, analyses, or charts.
- If your tables contain formulas or macros, create a second copy at the end of the project that contains only the results of those formulas.
- If possible, use only Latin (English) alphanumeric characters (a-z and 1-10) in data and headers. Avoid the use of commas in data if possible.
    - If your data includes non-alphanumeric characters, e.g., letters with diacritics (accents), always check your data to see that it has been correctly interpreted when you open or reuse the file in different software applications.
    - If your data must include commas, and you are saving the file as comma or tab delimited, make sure to qualify or "escape" the data between the columns by adding double quotes around the data values.

Rules for Rows

- A few files with many rows is preferable to many files with few rows. However, you may want to consider splitting files with more than 1,000,000 rows or 15,000 columns depending on what programs are typically used to read the data.
- Each row in your file should represent a single record or data point, e.g., the measurements of one sample or the response of one individual.
- The first row in the table should be reserved for column headers, aka field names.
    - Each column header should be concise but meaningful, contain only alphanumeric characters (with the addition of hyphens or underscores if necessary) and should never be duplicated in the same table.
    - If possible (and relevant), include units of measurement in the column header.

Data Standardization

- Standardize the format of data within each column, e.g., calculate numerals to a set decimal place.
- Use international, e.g., ISO; national, e.g., FGDC; or field-specific, e.g., LCSH; standards when collecting common types of data.
    - Use the ISO standard for recording dates – four digit year first, then two digit month, then day, e.g., 2018-01-31
- Decide on a consistent way to indicate missing data, and stick with that convention! Document that convention in your data dictionary (see Error! Reference source not found.)
    - Common ways to indicate missing data is to use a code such as -999 or -9999, or use text like "missing"
    - Always check to make sure that your missing data is interpreted correctly in any software you use to analyze or process it.
- Provide a data dictionary that explains the contents of your tabular files and gives additional context, including what any abbreviations mean, the units used, and any standards followed.

- The data dictionary should be named similarly to the data file (see file naming best practices). If you use Excel and want to keep the data dictionary as a separate "tab" in the file, that is acceptable, but be aware that other software applications may not be able to correctly interpret the relationship between the contents of the two tabs.

Examples

Fig. 1 Well-formatted tabular data

Site	Ecosystem	Plot	Depth_cm	Section_length_cm	Total_core_length_cm	Percent_LOI	Percent_TC
Al Aryam	salt marsh	6	30-50	20	-9999	5.818	10.26
Al Aryam	salt marsh	6	50-81	31	-9999	3.813	10.58
Eastern Mangrove	salt marsh	1	0-15	15	85	4.861	11.14

Fig. 2 Poorly formatted tabular data

						Data collated by Dr. R.E. Searcher 1/10	
Soil carbon data							
	Ecosystem	Plot #	Depth	Sect length	core	LOI	%TC
Al Aryam	salt marsh	6	30-50	20		5.82	10.26
Al Aryam	sm	6	50-81	31		3.813	10.58
Eastern Mangrove	salt marsh	1	0-15	15	85	4.861	11.14

References

2007. Best Practices for Preparing Environmental Data Sets to Share and Archive. Hook, L.A., Beaty, T.W., Santhana-Vannan, S., Baskaran, L., & Cook, R.B. http://daac.ornl.gov/PI/bestprac.html

2009. Some Simple Guidelines for Effective Data Management. Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. Bull. Ecol. Soc. Am. 90(2)205-214. http://www.nceas.ucsb.edu/files/news/ESAdatamng09.pdf

Preparing tabular data for description and archiving. Cornell University Research Data Management Service Group. https://data.research.cornell.edu/content/tabular-data

Expressing intentional blanks (null values) in a tabular dataset. DataOne. http://www.dataone.org/best-practices/identify-missing-values-and-define-missing-value-codes

2017. Ecology Tutorial: Data Organization in Spreadsheets. Data Carpentry. http://www.datacarpentry.org/spreadsheet-ecology-lesson/