RESEARCH DATA MANAGEMENT SERVICE GROUP
Comprehensive Data Management Planning & Services

(http://www.cornell.edu)

Search

Preparing tabular data for description and archiving

These are general guidelines for preparing tabular data for inclusion in a repository or for sharing it with other researchers, in order to maximize the likelihood of long-term preservation and potential for reuse. Individual repositories may have different or more specific guidelines than those presented here.

- General guidelines
- Data organization and formatting
- Data quality assurance
- Tools to help clean up tabular data

General guidelines

- Only include data in a data file; do not include figures or analyses.
- Consider aggregating data into fewer, larger files, rather than many small ones. It is more difficult and time consuming to manage many small files and easier to maintain consistency across data sets with fewer, larger files. It is also more convenient for other users to select a subset from a larger data file than it is to combine and process several smaller files. Very large files, however, may exceed the capacity of some software packages. Some examples of ways to aggregate files include by data type, site, time period, measurement platform, investigator, method, or instrument.
- It is sometimes desirable to aggregate or compress individual files to a single file using a compression utility, although the advisability of this practice varies depending on the intended destination repository.
- Individual repositories may have specific requirements regarding file formats. If a repository has no file format requirements, we recommend tab- or comma-delimited text (*.txt or *.csv) for tabular data. This maximizes the potential for use across different software packages, as well as prospects for long-term preservation.

Data organization and formatting

Organize tabular data into rows and columns. Each row represents a single record or data point, while columns contain information pertaining to that record. Each record or row in the data set should be uniquely identified by one or more columns in combination.

Tabular data should be "rectangular" with each row having the same number of columns and each column the same number of rows. Fill every cell that could contain data; this is less important for cells used for comments. For missing data, use the conventions described below.

Column headings

Column headings should be meaningful, but not overly long. All column headings within a file should be unique. Assume case-insensitivity when creating column headings. Use only alphanumeric characters, underscores, or hyphens in column headings. Some programs expect the first character to be a letter, so it is good practice to have column headings start with a letter. If possible, indicate units of measurement in the column headings and also specify measurement units in the metadata.

Use only the first row to identify a column heading. Data import utilities may not properly parse column headings that span more than one row.

Examples of good column headings:

max_temp_celsius - not max temp celsius (includes spaces, doesn't include units)
airport_faa_code - not airport/faa code (includes special characters)

Data values and formatting

- Use standard codes or names when possible. Examples include using Federal Information Processing (FIPS) codes (https://www.census.gov/library/reference/code-lists/ansi.html) for geographic entities and the Integrated Taxonomic Information System (ITIS) (http://www.itis.gov/) for authoritative species names.
- When using non-standard codes, an alternative to defining the codes in the metadata is to create a supplemental table with code definitions.
- Avoid using special characters, such as commas, semicolons, or tabs, in the data itself if the data file is in (or will be exported to) a delimited format.
- Do not rely on special formatting that is available in spreadsheet programs, such as Excel. These programs may automatically format any data entered into a cell, which can include removing leading zeros or reformatting date and time cells; in some cases, this may alter the meaning of the data. Some of these changes revert the cell back to its original value when changing the cell type to a literal 'text' value and some do not. Changing cell types from "General" to "Text" before initial data input can prevent unintended reformatting issues.

Special types of data - Date/Time

- Indicate date information in an appropriate machine-readable format, such as yyyymmdd or yyyy-mm-dd (yyyy: four-digit year; mm: two-digit month; dd: two-digit date). Indicate time zone (including daylight savings, if relevant) and use of 12-hour or 24-hour notation in the metadata.
- Alternatively, use the ISO standard for formatting date and time strings. The standard accommodates time zone information and uses 24-hour notation:yyyymmdd or yyyy-mm-dd for date; hh:mmTZD for time (hh: two-digit hour, in number of hours since midnight; mm: two-digit minutes; ss: two-digit seconds; TZD: time zone designator, in the form +hh:mm or -hh:mm, or Z to designate UTC, Coordinated Universal Time).

Special types of data - Missing data

- Use a standard method to identify missing data.
  - Do not use zeroes to represent missing data, and be cautious and consistent when leaving cells blank as this can easily be misinterpreted or cause processing errors.
  - Depending on the analysis software used, one alternative is to select a code to identify missing data. Common conventions include using:
    - -999 or -9999
    - NA, N/A, or NULL
- Indicate the code(s) for missing data in the metadata.
- When exporting data to another format, check to ensure that the missing data convention that you chose to use was consistently translated to the resulting file (e.g. be certain that blank cells were not inadvertently filled).

Data quality assurance

Consider performing basic data quality assurance to detect errors or inconsistencies in data. Here are some common techniques:

- Spot check some values in the data to ensure accuracy.
- If practical, consider entering data twice and comparing both versions to catch errors.
- Sort data by different fields to easily spot outliers and empty cells.
- Calculate summary statistics, or plot data to catch erroneous or extreme values.

Providing summary information about the data and including it in the metadata helps users verify they have an uncorrupted version of the data. This information might include number of columns; max, min, or mean of parameters in data; number of missing values; or total file size.

Tools to help clean up tabular data

OpenRefine (http://openrefine.org/) (formerly GoogleRefine) is a very useful tool for exploring, cleaning, editing, and transforming data. Advanced operations can be performed on data using GREL (OpenRefine Expression Language) (https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions).

References

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

Best Practices for Preparing Environmental Data Sets to Share and Archive (https://corpora.tika.apache.org/base/docs/govdocs1/602/602668.html). Hook, L.A., Beaty, T.W., Santhana-Vannan, S., Baskaran, L., & Cook, R.B. 2007.

Ecological Data: Design, Management and Processing. Michener, W.K. & Brunt, J.W. (Eds.). 2000.

Guide to Social Science Data Preparation and Archiving (http://www.icpsr.umich.edu/files/deposit/dataprep.pdf). Inter-university Consortium for Political and Social Research. 2009.

Some Simple Guidelines for Effective Data Management (https://esajournals.onlinelibrary.wiley.com/doi/full/10.1890/0012-9623-90.2.205). Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. Bull. Ecol. Soc. Am. 90(2)205-214. 2009.

Page last updated Apr. 2023


(http://www.twitter.com/curdmsg)