ORNL DAAC guidelines for CSV data files

Updated Feb 22, 2017

The ORNL DAAC will standardize our tabular data files in order to

1) improve data usability for both humans and tools, and
2) integrate with automated visualization and subsetting via a web feature service.
The following standards should guide QA of tabular data files.

- Do not rely on Excel, but open the file in a text editor for a final check before QA is complete.
- The QA staff may ask the investigator to correct the files, or do it themselves, depending on the effort required.
- Any exceptions to the standards should be clearly noted in the documentation.

General format requirements:

✓ Use ".csv" as the file extension
✓ File names in snake case i.e. lowercase with underscores: "example_file_name.csv"
✓ Header section consists of one row of column names followed by one row of units
✓ No ASCII control characters
✓ UTF-8 encoding is preferred, unless other encoding (e.g. UTF-16) is required.
✓ Use Unix-based newline/line ending/end of line (EOL) characters
✓ No DOI or URLs
✓ Columns delimited by commas
✓ No empty lines or rows, especially pay attention to the last line in a file
✓ Each row should have the same number of columns

For spatial files:

✓ Latitude and longitude in separate columns with names: "latitude" and "longitude"
✓ If a data file is not in WGS84 (EPSG:4326) CRS, use CoordX and CoordY as the column names
✓ Latitude and longitude in decimal degrees
✓ A CSVT file (with same filename but a .csvt extension) can be used to explicitly specify the data type of each column (e.g. "Integer", "String", "Real", "CoordX", and "CoordY"), especially if the data file will be ingested into SDAT. (See https://giswiki.hsr.ch/GeoCSV#CSVT_file_format_specification for further details)
✓ If coordinates are not in WGS84 (EPSG:4326) CRS, an optional *.prj file (same as the Shapefile format) can be used to explicitly specify the CRS.

Data content and formatting:

✓ No spaces in column names
✓ Column names in snake case. i.e. lowercase with underscores: "example_column_name"
✓ Missing data for numeric columns = -9999 or additional decimal 9's to an appropriate level of precision
✓ Missing data for text columns (any blank cell) = "NA"

✓ If a column value string contains characters, such as commas and quotes, place it in double quotes. If a column value string contains double quotes, double the double quotes to escape them and then place the value string inside a pair of double quotes. This can be accomplished when saving the file.

✓ Dates and times in 24 hour UTC. No local time zones unless UTC is also provided.

✓ Dates in YYYY-MM-DD format

✓ Times in hh:mm:ss (or hh:mm:ss+nn if time zone needs to be included) format

✓ Named sites and locations should have an associated geographic location

✓ Text and numeric data should not be mixed in the same column (see https://daac.ornl.gov/PI/BestPractices-2010.pdf), as they are in

estimated_depth	shrub_cover
4	30
6	> 75
4 to 5	25
5	65

References:

1. GDAL CSV Format Driver, http://www.gdal.org/drv_csv.html
2. GeoCSV Specification, https://giswiki.hsr.ch/GeoCSV