08/20/2012

Data Set Best Practices for EPSCoR Researchers

In an effort to streamline the data ingest and documentation process EDAC has created a collection of best practices for collecting information related to your data sets. These best practices are recommendations, some of which may not apply to your data set. We will continue to update this document as needed to help streamline the process of data collection. Thank you for your continued participation and patience in this ongoing process.

1. Denote missing data with a universally recognized code like -9999 that is outside the expected data range.
2. Parameters should be listed horizontally in column format with sites and or/date-time stamps listed vertically as rows. See below for an example.
3. If it is easier or more convenient for you, multiple sites can be combined into one data set provided that each site is represented by a unique code or name. Otherwise you can submit data sets from separate study sites as separate files.
4. Data should be submitted in the exact format that the researcher wishes to be published because EDAC personnel cannot modify anyone's original data. This can include but is not limited to adjusting significant digits, changing column names, etc.
5. Researchers should include a field for latitude, longitude, and elevation when applicable to provide the location of their study sites/locations.
6. Data submitted in a spreadsheet is typically in one of four types:
    a. Sensor Data (Range Domain) – For sensor data like temperature or dissolved oxygen, be sure to indicate the maximum and minimum values that the sensor can collect, the precision of the instrument (usually represented by a decimal value such as 0.1 or 0.001), as well as the units of measure. If you provide us with the brand and model of your sensor we can check the manufacturer's website before sending you email requests for clarification.
    b. Categorical/Qualitative Data (Enumerated Domain) – If you are using codes like good, fair, poor or x, ., ! to represent values then be sure to define each of those values. It is especially important to define qualitative terms like good, fair, or poor by providing us your definitions for those terms or the methods used to calculate those terms. Future researchers will not know what those codes or symbols represent without this information.
    c. Existing Code Values (Code Set Domain) – If you are using an established, pre-existing set of code values and definitions you only need to tell us what the code set is and who the source is. For instance, if you are using FIPS codes to denote census boundaries we only need to know FIPS codes and US Census Bureau.
    d. Free Text (Unrepresentable Domain) – Free text fields within spreadsheets are usually comment, site name, city, county, or state fields that contain text that does not fall into one of the other three categories.

    e. If your data does not seem to fit into one of these four categories please contact us and we will help make sure the data is documented properly.

7. If you have ideas for improving the data collection process, please feel free to contact us and let us know. We want to make this process as quick and painless for you as possible.
8. If you would like samples of data sets that employ these best practices that previous researchers have provided, we are happy to provide those.

This table illustrates a fictitious data set that employs the best practices EDAC recommends:

Date	site_id	Latitude	Longitude	Elevation	pH	Turbidity	Comments	pH_quality
2012/01/01	RRWWTP	35.33	-106.54	5245	3.29	24.9		Good
2012/01/02	RRWWTP	35.33	-106.54	5245	-9999	-9999	Sensor buried	Poor
2012/01/03	RRWWTP	35.33	-106.54	5245	5.95	25.5		Good
2012/01/01	RGCentral	35.42	-106.43	5240	5.67	26.8		Fair
2012/01/02	RGCentral	35.42	-106.43	5240	6.04	29.5		Excellent
2012/01/03	RGCentral	35.42	-106.43	5240	6.18	29.2		Excellent
2012/01/01	RGAlameda	34.51	-106.50	5237	6.21	35.4	Low battery	Poor

EDAC/EPSCoR Metadata Creation Team

Su Zhang - szhang@edac.unm.edu

Michael Camponovo – mcamponovo@edac.unm.edu