7 best practices for environmental data management

1. Assign descriptive (and unique) file names

File names should reflect the contents of the file and include enough information to uniquely identify the data file. File names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type. The file name should be provided in the documentation and in the first line of the header rows in the file itself.

2. Use consistent and stable file formats

In choosing a file format, data collectors should select a consistent format that can be read well into the future and is independent of changes in applications. Excel, as an example, is a useful tool for data manipulations and data visualization, but versions of Excel files may become obsolete and may not be easily readable over the longer term. Likewise, database files can be a very effective way to store and manipulate data, but the raw formats tend to change over time (even a few years). If your collection operation has used proprietary file formats, creating an export in a stable, well-documented, and non-proprietary format is important for maximizing others' abilities to use and build upon your data.

3. Define the content of your data files

The contents of the data files flow directly from experimental plans and are informed by the destination archive and data dissemination plans. Parameters and units and other coded values may be required to follow certain naming standards. In order for others to use your data, they must fully understand the contents of the data set, including the parameter names, units of measure, formats, and definitions of coded values such as flags and missing values. Provide the English language translation of any data values and descriptors (e.g., coded fields, variable classes, and GIS coverage attributes) that are in another language.

4. Use consistent data organisation

We recommend that you organize the data within a file in one of two ways. Whichever style you use, be sure to place each observation in a separate line (row). Most often each row in a file represents a complete record, and the columns represent all the parameters that make up the record. This arrangement is similar to a spreadsheet or matrix. A second arrangement may be more efficient when most records do not have measurements for most parameters, that is, a very sparse matrix of data, with many missing values. In this arrangement, one column is used to define the parameter and another column is used for the value of the parameter. Other columns may be used for site, date, treatment, units of measure, etc.

5. Perform basic quality assurance

In addition to scientific quality assurance, we suggest that you perform basic data QA on the data files prior to sharing. When QA is finished, describe the overall quality level of the data.

6. Assign descriptive data set titles

We recommend that data set titles be as descriptive as possible. The title may be the first thing people will see when looking at a dataset. So making descriptive titles is important for people searching for data. When giving titles to your data sets and associated documentation, please be aware that these data sets may be accessed many years in the future by people who will be unaware of the details of the project.

7. Provide documentation

In order for your data to be identified, found, accessed, and used properly by others in the future, they must be thoroughly documented. The documentation accompanying your data set should be written for a user 20 years into the future. Therefore, you should consider what that investigator needs to know to use your data. Write the documentation for a user who is unfamiliar with your project, sites, methods, or observations.