Smithsonian Data Management Best Practices

Naming and Organizing Files

Name and organize your files in a way that indicates their **contents** and specifies any **relationships** to other files.

The five precepts of file naming and organization:

- Have a **distinctive**, **human-readable** name that gives an indication of the content.
- Follow a **consistent pattern** that is machine-friendly. (see below)
- Organize files into **directories** (when necessary) that follow a consistent pattern.
- **Avoid reuse** of semantic elements in nested file and directory names that might cause confusion if files are removed from those directories.
- Have a **file extension** that matches the file format (no changing extensions!)

FILE NAMING

A file name should enable **disambiguation** among similar files and, for large numbers of files that make up a dataset, facilitate sorting and reviewing. Ideally, file names should be **unique**.

Keep in mind that files can be moved and, without the inherited folder structure, important descriptive information about the contents could be lost. Consider whether a filename would be meaningful outside of your chosen directory structure, and if not, how important the loss of that context would be, e.g., if the date a file was created is important, include it in the filename rather than just the directory name.

To provide a description of the file contents in the name itself, you should include elements such as:

- a **date**, or at least the year, the contents of the file were created, in the **YYYYMMDD** format (four digit year, two digit month, two digit day.)
    - start the filename with the date if it is important to store or sort files in chronological order.
- the **project name**, or documented abbreviation for the project.
- an **accession or other standard record number** if data is based on or includes SD-600 collections.
- your **organization's name** or abbreviation (if files are to be shared among collaborators.)
- the **location** related to the contents of the file, such as city, research site, etc.
- a **version number**, prefaced by "v", or another indicator of the file content's status such as "_draft" "_final" or similar.
    - Avoid *starting* the filename with version number, "draft" or "final"
- an **ordinal number padded with zeros** (particularly if the file needs to be sequenced/sorted with many other files).
    - use a minimum of two zeros for padding, with as many as necessary to accommodate the quantity of files you expect, e.g., if you expect 1,200 data files from one instrument, pad the filenames with three zeros, starting with _0001

Filenames for any given project or program should follow a **consistent pattern**. Adopt a pattern that will enable you to make filenames unique within each project, and are machine-friendly.

- Omit spaces and punctuation other than **hyphen** and **underscore.**
- Use **underscore** or "**camelCase**" between file name elements, e.g., my_data_file.txt or myDataFile.txt . Neither approach is better - just choose *one* and stick to it!
- Do not use spaces, tabs, semicolons or periods to separate elements of a filename.
- Try to use only **ASCII-encoded alphanumeric characters**, e.g., letters found in the Latin alphabet, and numbers between 1 and 10.

- Limit the name to **25 characters** in length if possible. **Short but meaningful is best.**

Examples of well-formed file names

1. For an image of a specimen in the Fishes collection, NMNH, collected in Mindoro, Philippines in 2000 with the catalog number USNM 379221 (3 options) :

    a. 2000_USNM_379221_01.tiff
    b. USNM_379221_01.tiff
    c. PHL2000USNM379221.tiff

2. A versioned file of tabular data and the accompanying data dictionary for a project in 2018 called “Multi-site cross-cutting longitudinal study” (two potential abbreviations for the project are given):

    a. 2018MSCCLSv1.txt ; 2018MSCCLSReadMe.txt
    b. MultisXxLong2018v1.txt ; MultisXxLongAbout.txt
    c. 2018_MSCCLS_v1.txt ; 2018_MSCCLS_readme.txt

Tip: You can bulk **rename** and manipulate files by scripting in the programming language of your choice, using PowerShell (Windows) or the Finder (Mac), or you can use an application like:

- Adobe Bridge
- Bulk Rename Utility: http://www.bulkrenameutility.co.uk/Main_Intro.php
- Renamer 5 for macOS: https://renamer.com/
- PSRenamer (requires JRE/JVM): http://www.powersurgepub.com/products/psrenamer/index.html

FILE ORGANIZATION

Like file naming, **consistency** is key. Organize files in a way that makes sense within the context of your project, but would also make sense to someone who was not intimately familiar with your project.

How files are nested in directories can be dependent on the **number of files** you are working with, and what aspect of those files is **most important for analyzing** or re-using the information in them.
For instance, if you have hundreds of thousands of image files collected over many years from many different locations, you may want to organize first by year, then month, then location. You could also organize them entirely by date, and include the location in the filename. Alternatively, organize by location, and only include the date in the filename.

If you are working on a collaborative project, make sure all collaborators are using the same principles to organize and name files!

Examples of Directory Organization

Example 1: SI and UCSD are both contributing to a five year project that involves taking measurements over time on two sample materials, A and B. Submitted files are for analyzed rather than raw data, and UCSD is employing two methodologies for analysis, submitted as versions.
Because the date of the measurement is important, files are first named by date, then sample. Each are organized into directories by contributor, and further grouped at the top level by year.

- 2017
    - UCSD
        - 20171001_B_v1.csv
        - 20171001_B_v2.csv
        - 20170930_B_v1.csv
        - 20170930_B_v2.csv

    - SI
        - 20170930_B_SI.csv
        - 20170925_A_SI.csv

Example 2: Images and corresponding description of those images from various sites in Pennsylvania, taken over the course of several years. The researcher expects to have between 150-300 images per site per year. In this example, a text file with descriptive metadata for all the images taken on one day is stored in a separate directory. This metadata file could also be co-located with the images.

- 2017_Images
    - Philadelphia
        - phil_20171028_001.tiff
        - phil_20171028_002.tiff
        - phil_20171028_003.tiff
        - phil_20171029_001.tiff
        - phil_20171029_002.tiff
    - Pittsburgh
        - pitt_20170922_001.tiff
        - pitt_20170922_002.tiff
        - pitt_20170922_003.tiff
- 2017_Metadata
    - Philadelphia
        - phil_20171028.txt
        - phil_20171029.txt

REFERENCES

Briney, Kristin. 2015. Data management for researchers: organize, maintain and share your data for research success.

Purdue Library. 2017. *Data Management for Undergraduate Researchers: File Naming Conventions*. http://guides.lib.purdue.edu/c.php?g=353013&p=2378293

Stanford Libraries. (viewed 2018). *Best practices for file naming*. http://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

University of Edinburgh. 2007. Records Management: Naming Conventions. https://www.ed.ac.uk/records-management/guidance/records/practical-guidance/naming-conventions