**Data Management Guide for Wild Acadia**

May 2022

**Introduction**

This data management guide is designed to highlight Schoodic Institute’s current recommendations for data storage, formatting, and preparation for ease of use (i.e., analysis, visualization, etc.). The information in this guide is meant to help streamline Wild Acadia’s interaction with data. This includes providing a consistent, functional approach for maintaining data, and creating datasets that are ready for modern analyses, as well as a guide for storing and archiving data to avoid loss of this important information.

This guide is organized into three sections, that in combination provide a baseline for data management. First, the *Directories, subdirectories, and files* section discusses what we have found to be the best naming conventions and organization to make sorting, finding, using, and analyzing the data easy and fast. Then, the *Data formatting* section goes into the optimal ways to format your data so they are best used by analysis software. Finally, the *Data archiving and storage suggestions* section outlines our recommendations for storing data and avoiding data loss.

**Directories, subdirectories, and files**

1. Directory organization and naming
   1. We suggest having a separate directory for each project, for which all data and data outputs for that project can be stored within.
   2. It is often better to divide large projects into subprojects to avoid overcomplication of the directory.
   3. The directory name should:
      1. use clear wording that could be understood by someone completely unfamiliar with your project,
      2. avoid special characters and capitalization (the only character that should be used in names is an underscore),
      3. and contain a year or month at the end if applicable/desired following standard ISO 8601 format.
   4. Each project directory should also include a metadata (we recommend an excel spreadsheet or google sheet) that:
      1. lists the filenames for all files within the project directory,
      2. lists the file path for each file,
      3. and contains a brief description of each file.
2. Subdirectory organization and naming
   1. We recommend that within each project directory, there be a “data”, and an “outputs” subdirectory.
   2. The data subdirectory should be named “data” and contain two additional subdirectories:
      1. “raw” - for all raw data files to be stored in (file path = “project\_directory/data/raw”),
      2. and “processed” - files that have been manipulated into the proper formatting for analyses from the raw file if necessary (file path = “project\_name/data/processed”).
   3. The outputs subdirectory should be named “outputs” and contain anything produced from the data (e.g., tables, figures, markdown reports, images, etc.).
   4. Additional subdirectories may be necessary depending on the goals and data for the project.
3. File naming conventions
   1. The basis of our suggested naming conventions can be summarized by three rules:
      1. use clear wording that could be understood by someone completely unfamiliar with your project while keeping it as short as possible,
      2. avoid special characters and capitalization (the only character that should be used in names is an underscore),
      3. and end filenames with the date in standard ISO 8601 format: “YYYYMMDD” (this makes sorting files by date very easy).
   2. Therefore, files would be named using this template:
      1. “filetitle\_breifdescription\_YYYYMMDD”
   3. As an example, a download of an iNaturalist dataset from Acadia National Park accessed on May 1st could be named:
      1. “inatuarlist\_acadia\_raw\_20220501"
   4. It is important to keep the file names short, while still maintaining enough information in the filename to identify it. This is where having a metadata for each directory becomes critical.

**Data formatting**

1. Typical datasheets
   1. In general, it is best practice to keep data in a flat file using Microsoft Excel or Google Sheets.
   2. These datasets should be formatted so that column A, row 1 is the first column name. Additional column names are to be listed in row 1 in their own cell in the following columns. The actual data values should then start in column A, row 2.
   3. Optimal column naming conventions are to:
      1. use lowercase letters,
      2. avoid numbers when possible,
      3. and use “.” or “\_” as separators, never use spaces or other characters (preferable “.” as it is easier to use during analysis).
   4. Leave missing values blank, do not fill with “na” or anything similar. Only enter “0” where the value is truly 0.
   5. Consistency is critical! Ensure consistent formatting of names, dates, and collected values. This is very important for any future use of the data to produce summaries, visualizations, reports, manuscripts, etc.
   6. Dates recorded can be in any consistent format that includes the day, month, and year. Time, if collected, should be recorded in a separate column in any consistent format.
   7. You should save versions of your data throughout the collection and entry periods so that you can always revert back to an older version. Save these files in the “raw” folder in the “data” subdirectory, and use the naming conventions specified above.
   8. When you have the full data entered in your spreadsheet, these raw data should be saved, stored, and archived as a “.csv” following the archiving and storage suggestions below.
   9. As a separate tab on your spreadsheet, you should include a metadata section that describes:
      1. when and where the data were collected,
      2. what project the data are associated with,
      3. who collected the data,
      4. and a description of each column
2. Other data
   1. In the case of data where it does not make sense to compile into a flat formatted spreadsheet (spatial, imagery, etc.), we recommend maintaining the default file type.
   2. In terms of images, “.png” is the preferred format.

**Data archiving and storage suggestions**

1. Local storage and archiving
   1. We recommend organization-owned external harddrives as a budget-friendly way to store copies of the data.
   2. Storing the entire project directory is ideal whenever possible, and compressing the file (“.zip”) will save space.
   3. If storing the entire project directory is not an option, focus on the raw data and the final processed data that are ready to be used are the priorities.
      1. These should be stored as “.csv” files.
2. Cloud storage and archiving
   1. We recommend storing all parts of the project directories on a cloud storage system, such as Google Drive, Dryad, or other cloud storage providers.
   2. If there are limitations and the full project directory cannot be stored, the raw data and the final processed data that are ready to be used are the priorities.
      1. These should be stored as “.csv” files.