iEcoLab DMP Quick Guide

Last Updated: 07 April 2021

Contents

"*Without data, you are just another person with an opinion.*" — W. Edwards Deming

"*Data is a precious thing and will last longer than the systems themselves.*" — Sir Tim Berners-Lee

---

Purpose

Data management is project management, and a data management plan (DMP) is used to standardize data formatting, naming, organization and storing so that data are accessible, understandable and reproducible. Data are any information collected or created pertaining to a research project (e.g. data tables, analysis scripts, figures, manuscripts, and reference libraries).

The definition of a research project depends on the individual project. Research projects are defined by project leads in collaboration with a PI(s). An easy guideline to follow is to use the data to define what is a project and what is a subproject. For instance, if you are using data that will be incorporated into many papers, then you would not want a project for each individual paper because then you would have multiple copies of data and keeping good data management practices across all copies would be difficult. In that case each paper would be a subproject under a larger project as defined by the data.

For the full Data Management plan visit WEBSITE

Updating

The DMP is reviewed regularly because each time a new project is started, the DMP should be referenced. It is updated based on input from iEcoLab team members who are actively managing and creating research projects. All iEcoLab members are encouraged to read this document and discuss with their PI(s) if they prefer to follow a different DMP for one project.

FurtherResources

- Borer, E.T., Seabloom, E.W., Jones, M.B. and Schildhauer, M., 2009. Some simple guidelines for effective data management. The Bulletin of the Ecological Society of America, 90(2), pp.205-214. LINK
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. and Teal, T.K., 2017. Good enough practices in scientific computing. PLoS computational biology, 13(6). LINK
- Guide to Reproducible Code
- Stanford Guide to the Best Data Management Practices

Pre-Project Planning

In this section details for what should be planned out before the start of a project.

Project Delineation:

- **Projects should be defined based on logical groupings and can include subprojects**

Project Roles:

- **Project roles need to be clearly defined for each project and subproject**
- **Project roles should include at least PI point of contact, project lead, and data manager**

Authorship:

- **Authorship guidelines should be made during pre-project planning**
- **Authorship guidelines need to be agreed upon by the PIs and Project Lead**

Data Access and Ownership:

- **Data should be kept on a Shared Drive since PIs and University own the data**
- **Data access rules need to be agreed upon before data collection begins**

Filesystem

In this section, guidelines for how to organize, name, and document your files in a directory are outlined.

File Organization:

- **Each project should have a separate folder with at least four subfolders:**
  - data
  - submission
  - references
  - R project folder(s)

File Naming:

- **File names should be easy to understand, give information about the data, and be consistent**

The iEco Lab file naming convention is:

$$< ProjectName > \_ < DataType > \_ < AuthorInitials > \_ < Version >$$

Naming Conventions for other file types are in the below table.

File Type	Convention
Base Naming Convention	$< ProjectName > \_ < DataType > \_ < AuthorInitials > \_ < Version >$
Audio-Visual Files	$< ProjectName > \_ < DataType > \_ < Date > \_ < Version >$
Spatial Data	$< ProjectName > \_ < DataType > \_ < Projection > \_ < Version >$
Manuscripts	$< ProjectName > \_ < ManuscriptName > \_ < AuthorInitials > \_ < Version >$
Edited Manuscripts	$< ProjectName > \_ < ManuscriptName > \_ < AuthorInitials > \_ < Version > \_ < EditorInitials >$

File Documentation:

- **Documentation files describe the data by stating their authors, methods, editing, and attributes**

- **The documentation files include Readme, Changelog, and Metadata files**
  - Readme - Describes the data
  - Changelog - Describes any changes you made to the various data versions
  - Metadata - Describes the data in each column of the data table
- **Documentation files should be save as text files**
  - The Metadata can be saved as a CSV

Naming Conventions for these types of files are:

Document Type	Naming Convention
Meta Data	$< DataFileName > \_METADATA$
Readme Files	$< ProjectName > \_(0)\_README$
Project Changelog Files	$< ProjectName > \_(0)\_CHANGELOG$
Individual Changelog Files	$< DataFileName > \_CHANGELOG$
Individual Readme Files	$< DataFileName > \_README$

Version Control:

- **Data version control should be done manually by adding the version at the end of the filename**
- **At least three versions of the data are required: raw, v0, and final**

Data

Data Collection:

In this section, guidelines for how field and laboratory data should be collected are described. These guidelines are not for the actual methods and techniques of generating the data but rather how to get data from the field to a data file on a computer.

Field Data Collection:

- **Data should be recorded with dark pen and never erased or scratched out**

Lab Data Collection

- **Data collected in the lab should kept together in a lab notebook and recorded in dark ink**

Key points in keeping a Lab Notebook:

1. Neat and legible handwriting in dark ink; not pencil if able
2. Procedure/Study title and purpose clearly stated
3. Methods described clearly and succinctly, with errors and steps taken to correct them
4. Calculations performed neatly showing intermediate steps
5. Errors crossed out with a single line, initialed, and briefly explained
6. All pages dated at the top and numbered at the bottom

Data Transcription

- **Digitize and back up data within a week after collection**
- **If multiple people are transcribing data, then they should each be working with separate spreadsheets**
  - The different spreadsheets are then compiled by the Data Manager

Quality Control During Transcription

- **Spreadsheets should resemble field data sheets as much as possible**
  - The spreadsheets are then formatted by the Data Manager
- **You can use column rules or data entry forms for transcription quality control**
- **A secondary check by someone who did not enter the data should be done**

Data Mining

- **Work from master source list and compile and backup regularly**

Data Formatting

- **Data files should be checked with open source software before publishing**

Data Tables

- **Data tables should be saved as CSV files and be in long format**
- **Column headers are required, should be consistent, and easy to understand (see iEco Header Conventions)**
- **All text should be formatted consistently with dates in the YYYY-MM-DD format**
- **All cells should have values and if a value is missing then an NA should be entered**

Spatial Data

- **Spatial data should be a CSV, shapefile, GeoTIFF, or netCDF file depending on the data type**

Audio-Visual Data

- **Avoid proprietary formats and include data information in the recording or picture file**

Data Backup:

- **Data should be backed-up weekly using the 3-2-1 strategy:**
  - **3 copies of your data**
  - **2 copies stored locally on different devices**
  - **1 copy stored off site (i.e. on the cloud)**
- **Data should ideally be backed up weekly and at least monthly**

Final Data Storage and Sharing

- **Final data storage should be on an online data repository**
- **Data sharing permissions is determined by the Project Lead and the PI(s)**

Reference Libraries

- **Reference libraries should be created through Zotero. Contact Dr. Helmus for access**

If you cannot use Zotero then you should add a 'references' folder and follow the following naming convention.

$$< Author(s) > \_ < Year > [< Brief Description >]$$

Project Packaging

When organizing your analyses for a research project, it's important to maximize two aspects: reproducibility and ease of sharing. This will make the life of your collaborators, and crucially the life of "future you", much easier. R packages provide a great tool to create collaborative and reproducible code. An R package bundles together the data your project requires, custom-made functions that can be used throughout your code, vignettes that describe your analyses step-by-step, and a thorough documentation that helps you and others to reproduce your analyses in the future. In practice, creating an R package boils down to 1) organizing your project folders and files in a clever and consistent way, and 2) documenting thoroughly your steps and the tools you use.

Guidelines for creating a project package can be found HERE
