PIE LTER Data and Metadata Best Practices

The goal of the PIE LTER data and information system is to provide a centralized network of information and data related to PIE. This network provides researchers access to common information and data in addition to protected long-term storage. Data and information are also easily accessible to local, regional, and state partners and the broader scientific community. Researchers associated with PIE are committed to the integrity of the information and databases resulting from the research.

Data management and design of research projects is coordinated through the information management team. Several meetings each year provide researchers the opportunity to communicate with the information management team regarding the design of the specific research project and subsequent incorporation of data and information into the EDI database. For immediate assistance, please contact the Information Manager at pie_im@mbl.edu.

Individual researchers are responsible for providing data and associated metadata in compliance with the LTER Network Data Access Policy. Researchers using PIE facilities are expected to comply with the LTER policy even if they are not funded by the LTER.

Data files should be submitted with the completed Metadata Template to the PIE LTER Information Manager (IM). The IM reserves the right to edit metadata content for data compatibility with PIE and the LTER Network. Individual researchers are responsible for quality assurance, quality control, data entry, validation, and analysis for their respective projects.

Resources

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10). https://doi.org/10.18637/jss.v059.i10

Whitlock, M.C. (2011). Data archiving in ecology and evolution: best practices. *Trends in Ecology and Evolution, 26*(2). https://doi.org/10.1016/j.tree.2010.11.006

EDI Resources for Data Authors

EDI Quality Assurance

EDI Cleaning Data and Quality Control

Data Submission Checklist

- ✓ Is the data quality checked?
- ✓ Do all data points have a unique identifier?
- ✓ Do all variables have descriptive and concise names?
- ✓ Is there a date column? Is it formatted as YYYY-MM-DD?
- ✓ Do all the data of one kind have the same number of significant digits (decimal points)?
- ✓ Are empty cells filled with a missing value code? Is it consistent throughout the file(s)?
- ✓ Are all data within the acceptable range for that data type?
- ✓ Are any outliers or questionable data marked in a flagging or comment column?
- ✓ Is there any punctuation besides underscores in the column headers?
- ✓ Are all the data in one column in a consistent format? (All numerical, all text, all datetime, etc.)
- ✓ Is the Metadata Template completed?

Tips for collecting quality data

Quality assurance
Data quality starts at the moment of sample collection/generation. Quality assurance (QA) is the set of steps taken to ensure data collection is developed and adhered to in a way that minimizes inaccuracies. The purpose of QA is to produce high-quality data while minimizing the need for later corrections.

Some examples of QA measures:

- **Calibration.** Is the instrument operating within specification?
- **Operational conditions.** Is the instrument operating within conditions for which it wasn't designed? (e.g. excessive heat)
- **Replication and redundancy.** Facilitates estimation of error and can reduce data gaps (e.g. sample and hardware replication)
- **Anti-fouling.** Minimize data drift (e.g. wipers, manual cleaning)
- **Real-time checks.** Receive alerts when incoming data are out of bounds so adjustments can be made.
- **Schedule and field log.** Remember key events.

Give each sample a unique identifier. We recommend doing this by labelling each sample with date, site, treatment, and replicate so that no two samples have the exact same identifying information. For example, if you collect three samples on the same date, from the same site and treatment group, each should have a unique replicate number or letter to differentiate between them.

Related to QA is the documentation of data collection and analysis, or metadata. Metadata encompasses “who, what, when, where, and why.” You should start recording metadata while you are collecting and analyzing data. We recommend using field and lab notes to record important information. When in doubt, write everything down!

Some examples of metadata to record during data collection:

- **Date and time.** Include time zone and daylight savings.
- **Location** (if collecting field samples). Use a designated site name or GPS coordinates.
- **Who** is collecting or analyzing the data.
- **Sampling methodology.** List Standardized protocols while noting any potential variation.
- **Overview of data collection infrastructure.** Only the parts that may affect data values need to be recorded.
- **Instrumentation.** Make, mode, accuracy, precision.
- **Quality assurance.** Measures implemented and perhaps rationale.
- **Deviations** from collection protocols and plans.

Quality control

Cleaning data and quality control (QC) occurs after data collection and is a large and sometimes subjective topic. However, there are some general guidelines for making data more usable. Deciding what data are good, like dealing with perceived outliers or errors in the numbers, should be done carefully and in a way that preserves the original values rather than overwriting them.

Transfer your data and metadata to a digital format as soon as possible while things are still fresh in your mind and keep copies of your field and lab notes. It's a good idea to check your data immediately to identify statistical outliers, collection errors, entry or copy/paste errors, or numbers outside an acceptable range. Ask yourself if the data makes sense and if they are within normal range for the parameter. Do they make sense in the context of related variables? Graphing can also be very useful for quickly spotting unusual data points. Be careful when using copy/paste to make sure that you're not pasting formulas or losing the identifiers for the data.

Value checking is implemented as tests designed to address issues likely to be present in the collected data. Some examples:

- **Duplicate records**. Measurement listed twice. Not a replicate measurement.
- **Sequential records**. Some data should be in a sequential order (e.g. dates and times).
- **Range**. Data out of range may indicate a faulty measurement (e.g. relative humidity 0 - 100%).
- **Persistence**. Constant values may indicate a faulty measurements
- **Slope change and steps**. In time series, these may represent instrument drift.
- **Internal consistency**. Data values fall within an established range at a sampling location (e.g. a list of species expected at a site)
- **Paired consistency.** Duplicate observers/instruments produce similar values and trends.

Data flags are useful in communicating value specific information from quality control results (e.g. a value is below a detection limit or is questionable). Data flags can be added to a table as new columns using the naming convention <variable>_flag (e.g. temperature_flag). For consistency, consider making a comprehensive set of codes to be used across all the data created by a project. All flagging codes must be defined in the metadata.

Everyone who's name is associated with the dataset is responsible for making sure the data and metadata are of good quality, so send your dataset around to your co-authors to check over!

How to prepare data for submission

Data table creation
Consider how another person would use your data and what they would need to know about it. Make sure you include:

- Date and time
- Site
- Treatment (if applicable)
- Plot (if applicable)
- Replicate (if applicable)
- Species (if applicable)

In addition, make sure that:

- There are no punctuation or symbols in the column headers besides underscores. (see: **File and variable naming**)
- There are no units in the column headers. Units are described in the metadata. (see: **File and variable naming**)
- The data within each column is consistent. If they're numerical, they should have the same number of decimal points and no characters besides missing value codes. If they're categorical, codes should be defined in the metadata. The date is in the format YYYY-MM-DD or another recommended format. (see: **Variable types**)
- Missing data has an appropriate missing value code (we recommend "NA"). This code is defined in the metadata. (see: **Missing values**)
- All equations and links should have been removed. If you are copy/pasting, you can do a special copy/paste as values only in Excel.

File and variable naming
When naming files and variables, be descriptive and concise. Don't include information that will be conveyed in the metadata. Names of files and variables should be constructed from entirely alphanumerics and underscores for the sake of machine readability. Spaces and symbols can be inconsistently represented across computational applications.

Bad file name: **soil properties 2010-2020.csv**

Good file name: **soil_properties.csv**

Bad variable name: **dissolved oxygen % saturation**

Good variable name: **dosat**

Variable types
Consistency within variables facilitates type wise operations allowing similar types of data to be combined and operated on together. Data become much more difficult to

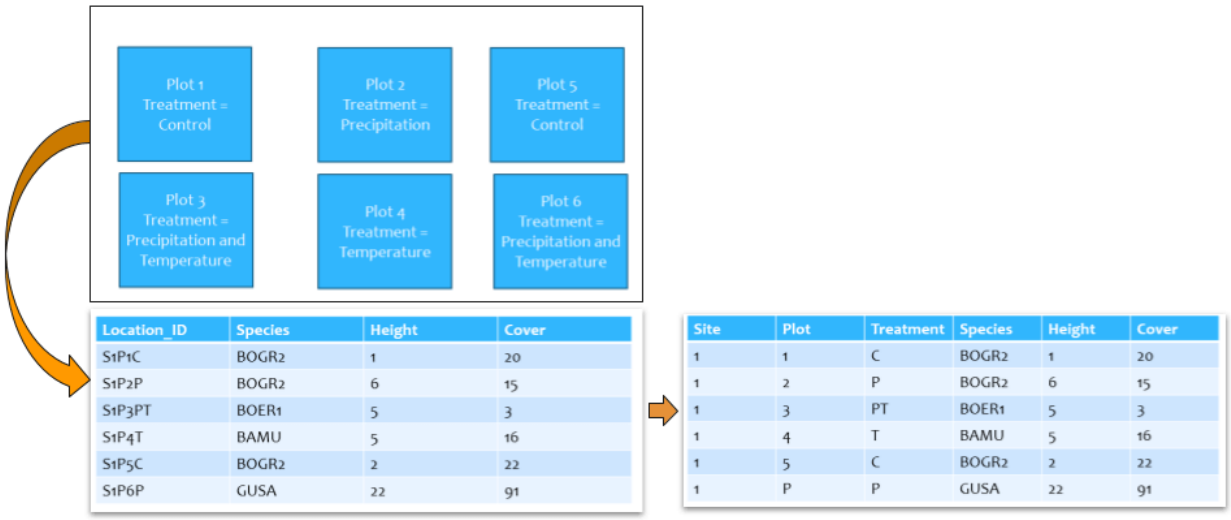
understand and use when variable types are mixed (e.g. numeric data mixed with character strings).

- **Dates and times.** The EDI Data Repository recommends the ISO 8601 Standard whenever possible (e.g. YYYY-MM-DD hh:mm:ss). Remember to also specify the time zone and daylight savings observation practices during measurements.
- **Numeric.** Numeric types should be consistent within a column (e.g. integer, real). If the measurements have a practical precision, the values within the column should be consistently represented in this precision, (i.e. keep meaningful numbers of decimals).
- **Categorical.** Categorical variables are frequently used for grouping data (e.g., experimental manipulation vs. control). Check for consistent representation in terms of spelling, abbreviations, casing, synonyms, etc. Provide definitions for each categorical value.
- **Character.** Character types should only be used when the other types don't apply. Numeric values should not be character type unless, possibly, the values represent identifiers and should not be used for calculations.

One value per cell

Within a column each cell should contain only one piece of information in a consistent format (e.g. datetime, numeric). The most frequent problems are comments entered into an otherwise numeric column (e.g. to denote a missing value). In that case it is recommended to have the value column and add a comment or flagging column where the text comments may be entered.

Another recommendation is to avoid overloading a cell with composite information. Below is an example illustrating the issue of more than one piece of information per cell. In the first table, the Location_ID column is a composite of multiple variables. Selecting values from a variable requires parsing. The second table follows the best practice of one piece of information per cell, where the data are easily accessed.



Location_ID	Species	Height	Cover
S1P1C	BOGR2	1	20
S1P2P	BOGR2	6	15
S1P3PT	BOER1	5	3
S1P4T	BAMU	5	16
S1P5C	BOGR2	2	22
S1P6P	GUSA	22	91

Site	Plot	Treatment	Species	Height	Cover
1	1	C	BOGR2	1	20
1	2	P	BOGR2	6	15
1	3	PT	BOER1	5	3
1	4	T	BAMU	5	16
1	5	C	BOGR2	2	22
1	P	P	GUSA	22	91

Missing values

Missing value codes denote when no observation was made. This differs from when data were collected and their quantity is zero or below detection. Applying consistent missing value codes within a table greatly simplifies reading into a software application. Empty cells should be filled with a missing value code to prevent software applications from interpreting these values differently, and to enable description within EML metadata. Commonly used missing value codes are "NaN" and "NA". However, every analytical software has its own preferred empty cell code.

Data table structure

We highly encourage the use of a “tidy dataset” structure, also known as a "long" table format. This means that each variable is a column and each observation is a row (Wickham 2014). If you need assistance structuring your dataset, please email the IM.

“Wide” format

Date	WS1	WS2	WS3	WS4
1/1/2001	1.1	5.4	0	9.1
1/2/2001	2.3	1.1	0	0.5
1/3/2001	0	2.4	4.1	3.2
1/4/2001	0	6.5	6.2	6.1
1/5/2001	4.5	0	0	9
1/6/2001	12	0	0	3.4
1/7/2001	0	0	0	6
1/8/2001	0	0	0	5.4
1/9/2001	1	12.3	0	4.3

“Long”, archive-ready format

Date	Site	Precip
1/1/2001	WS1	1.1
1/2/2001	WS1	2.3
1/3/2001	WS1	0
1/4/2001	WS1	0
1/5/2001	WS1	4.5
1/6/2001	WS1	12
1/7/2001	WS1	0
1/8/2001	WS1	0
1/9/2001	WS1	1
1/1/2001	WS2	5.4
1/2/2001	WS2	1.1
1/3/2001	WS2	2.4
1/4/2001	WS2	6.5
1/5/2001	WS2	0
1/6/2001	WS2	0
1/7/2001	WS2	0
1/8/2001	WS2	0
1/9/2001	WS2	12.3
1/1/2001	WS3	0
1/2/2001	WS3	0
1/3/2001	WS3	4.1
1/4/2001	WS3	6.2
1/5/2001	WS3	0
1/6/2001	WS3	0
1/7/2001	WS3	0
1/8/2001	WS3	0
1/9/2001	WS3	0

Data table aggregation

Data can be published as a set of tables or in aggregate as a single table. For instance, time series data could be split into files by year or by variable. Consider who will likely be using the data in the future and what format will be simplest for them to access, understand, and work with. If the data are split, try to maintain a consistent format so the tables can either be joined by key columns or "stacked" together.

Creating metadata

The Metadata Template goes into detail about the information you need to include, but in general you will need:

- A descriptive and specific title (it should be similar in format to the title of a published paper).
- Descriptions of the data table including each attribute (column), all codes used (including missing value codes), and the units for each attribute.
- The names, contact information, and ORCID (recommended) for each creator.
- A detailed abstract which provides enough context that users can fully understand the data.
- Keywords. We recommend starting with the LTER controlled vocabulary in order to improve future discovery and reuse of your data. Consider which keywords you would use to search for a similar dataset.
- Where the data was collected (in decimal latitude and longitude).
- When the data was collected.
- Which species were studied.
- Specific methods with enough information for another user to fully understand the data and replicate the experiment.
- Project information (PI name, e-mail, ORCID, and funding details).

If your dataset is an update to an existing dataset on EDI, please download the data package from EDI, make necessary edits, and send to the IM instead of creating a new data file and metadata file from scratch.