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.

Tips for collecting quality data

Data quality starts at the moment of sample collection/generation. You should start recording metadata while you are collecting and analyzing data. Metadata encompasses "who, what, when, where, and why." We recommend using field and lab notes to record important information such as date, time, location, what you are collecting and how, who's collecting the data, and what the project is. Other useful notes might include weather conditions, seasonal indicators, and anything out of the ordinary which might explain unusual data. When in doubt, write everything down!

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.

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.

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

We highly encourage the use of a "tidy dataset" structure. 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.

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.
- There are no units in the column headers. Units are described in the metadata.
- 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.
- 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.
- Missing data has an appropriate missing value code (we recommend "NA"). This code is defined in the metadata.
- The date is in the format YYYY-MM-DD or another recommended format.

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.

Data Checklist

○ Do all data points have a unique identifier?
○ Is there a date column? Is it formatted as YYYY-MM-DD?
○ Are all the data of one kind in the same units?
○ Do all the data of one kind have the same number of significant digits (decimal points)?
○ Are there any empty cells within the data file?
○ Do you have a missing value code? Is it consistent throughout the file(s)?
○ Are there any outliers?
○ Are all data within the acceptable range for that data type?
○ 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.)

References

Wickham, H. . (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

EDI Resources for Data Authors

EDI Cleaning Data and Quality Control