Data Submission Instructions

For LTER researchers

An T. Nguyen and Tim Whiteaker – BLE LTER Information Managers

2020-02-21

Orientation

LTER-funded data are required to be publicly accessible within two years of collection. We, your information managers (IMs), will help you meet that deadline by making sure your data are properly described with metadata, and by submitting the data and metadata (a "data package") to an appropriate archive for you. Contact us if you'd like to discuss how to organize your field data into meaningful datasets, and see this article from the Environmental Data Initiative.

As and even before you collect data, there are several helpful things to keep in mind:

- Remember to discuss what sort of data you will produce, when they will be ready, and when we need to archive them with the IM team. Always keep scans and physical copies of your field records; let us know if you'd like help organizing this task.
- Record dates and times consistently.
  - We prefer that date and time components are combined into one column; e.g. one "date_time" column instead of separate year, month, day, hour, minute, etc.
  - We strongly prefer ISO 8601 format "T" for date times. E.g. "2020-02-21T12:00-05:00". The letter T denotes separation between date and time. We have found using this ISO format prevents Excel from doing that annoying thing where it auto-changes date formats.
- Record latitude and longitude to at least six decimal places, and make note of the datum used, e.g., WGS 1984.
- If for any reason data are missing, i.e., "blank" Excel cells, note the cause(s).
- This article is an excellent resource on data organization in spreadsheets. Read it once and it'll pay off over the course of your research career, and make this process a pleasant & productive one!

The archiving and publishing process

1. Let us know you'd like to submit data. Obtain these from us: (1) Excel metadata template, (2) PDF instructions – this document, and (3) a zipped sample metadata+data package with example content. If you are reading this because you downloaded a zip from our website, the zip contains everything listed above.
2. Fill out the metadata template. Email the data and metadata to us.
3. Make sure you or someone familiar with your data is available to answer questions we may have about the data and metadata until the data package is officially archived.
4. Once the dataset is archived and has a DOI, please cite the data package in any related publications, both in a "Data Dissemination" section and in the references section. You might want to use the suggested citation available on the data archive's landing page.

What to submit

1. Your data file(s).

- We prefer quality-controlled raw data. Calculated or derived data (e.g., to make figures) are less useful to the end user. Exception example: derived oceanography
- For tabular data, organize in tidy format: one table per sheet, one header row, one row per observation, one column per variable, one value per cell. Also aim for every observation being spatially and temporally situated, i.e., every table has lat/lon and date-time columns.

2. A copy of the metadata template, completed as best you can.
3. Two Word documents containing dataset abstract and methods.
4. We would appreciate any publications or pre-prints that use the dataset to give us context for the metadata.

You may find the attached example submission package useful as you prepare your data. (Example courtesy of BLE-LTER's very own Vanessa Lougheed who patiently helped us refine the process. Thank you Vanessa!)

Metadata Template Instructions

The accompanying metadata template is an Excel workbook, oriented toward the Ecological Metadata Language, the metadata standard for the LTER network and widely used in the earth and environmental sciences. Refer to these instructions as well as the sample submission package, which will walk you through completing each sheet in the template workbook.

About the template workbook:

- Some columns contain tooltips. Click on the second through fourth rows to see them.
- Some columns contain built-in validation rules. In most cases this means you need to choose from a drop-down menu (hint: a downward arrow appears to the right of the cell). In some cases you will be able to enter a value not in the menu. Try to avoid pasting into columns with validation rules, as doing so might override them.
- You may notice some hidden columns; no need to worry about them.

Do *not* use superscripts or subscripts anywhere. Use ASCII characters – numbers and Latin letters with a few special symbols – most everywhere. Use underscores in file and column names, no spaces, no special characters. You may use a broader set of characters in abstract and method documents, such as the special character ġ in Utqiaġvik.

Dataset

Dataset title

- Ideally the title should be under 20 words. We might edit for clarity and brevity.
- The dataset title is distinct from publication titles. It needs to include the broad scientific theme, as well as some geographical, temporal, and taxonomic (if applicable) information about dataset.

Guidelines to abstract and method documents

- Send us two Word documents (or Google Docs; no Mac-specific formats please). In the template, write the exact file name, with extension. We generally use these documents as-is, so consider reviewing them for grammar and clarity before sending them.
- Spell out acronyms, limit the lingo. Use references if applicable (e.g., if you use an industry-standard protocol, but be explicit about any modifications). Any citation format is ok.
- The dataset abstract is distinct from publication abstracts. It provides more context for the dataset title: high level summaries of project goals, scientific questions, methodologies, format and scale of data, geographical, temporal, and taxonomical information. This abstract is ideally 100 words minimum according to DataONE's FAIR standards, but this is not a hard requirement.
- Your method document should describe how all of the information the dataset contains were obtained. Be explicit about equipment (models, calibration, etc.), and/or software/scripts used. Any processing or QA/QC performed on the data should be described.

Other columns

- Begin Date/End Date: use YYYY-MM-DD or YYYY.
- Lat/Lon Datum: we will assume this applies to all location coordinates in the dataset.
- Taxonomic Authority: specify if applicable to dataset.
- License: CC0 by default. Find out more at Creative Commons. While this license effectively places the data package in the public domain, no rights reserved, we will remind data users that it is both ethical and professional etiquette to give proper credit when using this dataset.
- Time Zone: time zone of date time values appearing in the dataset. Defaults to AKDT. Note that we've found some BLE sensors actually record in Alaskan standard time, which does not adjust for Daylight Savings Time and is equal to UTC/GMT -08. If your data follows DST, then there might not be a record on 2019-03-10 02:00:00 or from 2am to 3am, since DST skips ahead then.
- Update Frequency: is this a one-off dataset, or do you expect to add more data on a regular basis?

Personnel

List all personnel that should be given credit on the collection, curation, and maintenance of the data, after discussion with your team.

Role

The only required role is "creator," or people who have intellectually contributed to the dataset. Those listed as creators will appear as "authors" of the dataset and in auto-generated citations in most metadata display systems. Note that BLE funded datasets will list BLE LTER as the first creator per LTER's and our group's practices. There is no LTER controlled vocabulary for other roles. We recommend giving fair credit, and suggest possible roles below; you are welcome to credit contributors under other roles as you see fit.

Role	Who fits this role
creator	Intellectual contributors to the creation and/or continuation of dataset.
field technician field assistant	
data analyst	Person who processes and/or analyzes data; person who runs scripts, if distinct from programmer.
data entry lab technician lab assistant	
programmer	Writer/maintainer of scripts used to either collect, process, analyze, or maintain dataset.

Note for Core Program datasets: we have decided to list a singular creator on Core Program datasets "Beaufort Lagoon Ecosystems LTER, Core Program". All personnel involved will be listed under other defined roles. These roles and their duration will be recorded in a separate data table to be packaged and archived with the data.

Other columns

- Creator order: Per LTER best practices, the LTER site will be listed as first creator of the dataset. For each creator, list the preferred authorship order, starting with 2. For other personnel, leave blank.
- Fill in your organization, address, contact info, etc. *only* if this is your first time submitting *or* if there has been a change in contact information, e.g. moving to another institution. Otherwise, we most likely have your information already. Do list this information for non-officially BLE-affiliated personnel as we might not have it.

- We recommend including ORCIDs. ORCID is a persistent identifier system for researchers. You might want to encourage your team members to get one and fill out a profile; merely having an ORCID is no good without a profile.

Keywords

Keywords make your data easier to discover and provide context. However, keywords in controlled vocabularies – also known as keyword thesauri – serve these purposes better than idiosyncratic ones. Look for possible keywords in the thesauri linked in the template. You might want to re-use the keywords found in related publications. We will also add several keywords recommended by LTER best practices.

We recommend choosing at least one LTER Core Research Area, e.g., Primary Production.

Sites

Include lat/lon coordinates in decimal degrees (to at least six decimal places) for all sampling sites in your data, with a brief description of the site. While all supplied coordinates will be used in data tables to spatially situate each observation, we will list a smaller subset in the metadata to give data users an idea of the extent of sampling. Note that for Core Program stations, we often already have the coordinates.

Entities

A dataset can contain one or many data "entities," e.g., one table containing data from a group of similar surveys. If all your data are tabular, you might want to submit a single Excel workbook, in which case each "tab" or worksheet will be a data entity. If you are not sure how to organize your data into entities, or are submitting other types of data, let us know and we will discuss what constitutes a data entity in your case.

Attributes

Here we annotate each attribute in your dataset. Refer to the template tooltips as you go. Note that for tabular data, we use columns and attributes interchangeably, provided that the data follow tidy format. For non-tabular data, columns may not be an applicable concept, while attributes may still apply. While the above worksheets contain information that is applicable to the whole dataset, attributes are specific to each entity, so repeat attributes if the same one appears in two or more entities, and be sure to specify if the attribute differs in some way in different entities.

Attribute Codes

Codes apply to categorical variables, or attributes that allow a specific set of values. For example: you have a "WaterColumnPosition" column in a table, which denotes one of three possible sampling depths: "surface," "mid-column," and "bottom." In the template sheet, define all codes used in each categorical variable. Code-definition pairs are specific to each attribute, so repeat if you reuse codes in different attributes. Make sure to be consistent with codes in your data (e.g., don't use "M," "m," and "male" to refer to the same category).

Attribute Missing Values

> ***Leave No Cells Behind:*** *Never, ever, leave data blank.*

Always use code(s) to denote missing values in data, instead of leaving them blank. In the template, for each column in your data that has missing data, specify at least one missing code; you may skip columns without any. Make sure all codes entered in the template are used correspondingly in data, i.e. don't say the code is "not applicable" but have "NA" values in the spreadsheet. Code-definition pairs are specific to each attribute, so repeat if you reuse codes in different attributes.

Why use multiple missing value codes: The correct interpretation for missing values is scientifically important. For the same variable, data missing because "equipment malfunctioned" is very different from data missing

due to the observation being "under detection limit." We recommend keeping good field notes and not hesitating to use multiple codes to denote different underlying causes of missing data. See below for some possible causes of missing data, but enter your own as you see fit. For the following examples, suppose we are recording a set of traits on individual plants at the end of the growing season, including maximum fruit size.

Cause for missing data	Example
attribute under detection limit of equipment/method	Fruit smaller than ruler's smallest unit.
attribute over detection limit of equipment/method	Fruit larger than ruler.
attribute not applicable to this observation	Fruit size not an applicable concept to conifer sample.
attribute not available in this observation	Plant has yielded no fruit, although others in the same species have.
attribute not recorded due to external conditions	Observer forgot ruler at home, cannot measure fruit.