Data submission instructions for CAP LTER researchers*

CAP LTER Information Manager

2024-04-11

Table of contents

overview

LTER-funded research requires that data are made publicly available within two years of collection. The CAP LTER Information Manager will help you to meet that deadline by making sure your data are formatted intuitively, described properly with metadata, and by submitting the data and metadata (a "data package") to an appropriate data repository. Contact the CAP LTER Information Manager if you would like to discuss how to organize your data into logical datasets.

The Environmental Data Initiative provides a good overview of the data publishing process.

*thanks to the BLE and MSP LTERs for source materials and inspiration

The language in these instructions and associated metadata templates is oriented to CAP LTER data submissions, but most elements are generic and researchers providing data not associated with CAP LTER research can safely ignore references to the LTER and CAP LTER.

the archiving and publishing process

1. Obtain the template metadata files from this Google Drive directory. You can either make copies of the templates in Google Drive or download them (individually or as a zipped directory), however you prefer to work is fine.
2. Fill out the metadata templates: metadata_template (Google Sheet/Excel workbook), abstract (Google Doc/text file), methods (Google Doc/text file), and email to or share with the CAP LTER Information Manager the completed forms along with the data files.
3. The CAP LTER Information Manager will reach out with questions, and will provide an unpublished draft of the dataset prior to making it public. At this point, you would check that all of the documentation is correct. Work with the CAP LTER Information Manager to make corrections if needed then confirm your approval. Once approved, the CAP LTER Information Manager will provide a DOI to the published dataset for your records.
4. Once archived, please cite the data package in any related publications both in the data availability statement, if relevant, and, importantly, in the references section of the publication. Use the suggested citation available on the data archive's landing page.

what to submit

1. Your data file(s).

- All data should be quality-controlled, raw data.
- CSV is the preferred (but not required) format for tabular data.
- Tabular data should be organized in tidy format.
- GeoJSON is the preferred (but not required) format for vector data.
- GeoTIFF is the preferred (but not required) format for raster data.

2. The completed metadata templates:

- metadata_template (Google Sheet/Excel workbook)
- abstract (Google Doc/text file)
- methods (Google Doc/text file)

metadata template instructions

The metadata template is a Google Sheet/Excel workbook oriented toward the Ecological Metadata Language, the metadata standard for the LTER network and widely used in the earth and environmental sciences.

About the template workbook:

- Some columns contain tooltips.
- 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.

Do not use superscripts or subscripts anywhere. Use only ASCII characters (numbers and Latin letters with a few special symbols). As much as possible, use underscores in file and column names, avoiding spaces and special characters. You can, however, use Markdown formatting in the abstract and methods.

tab: dataset

data package title

- 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 the dataset.
- A dataset title is analogous to a journal article title: it should be descriptive, information rich, and crafted to provide readers with a sense of whether the resource is relevant to their interests.

geographic coverage description A general overview of the location where the data were collected. Default text for the CAP LTER study area is provided but edit as appropriate. If more detailed geographic information is relevant, those should be provided (also) as a data entity in tabular and/or geographical format (e.g., GeoJSON).

study area bounding coordinates Provide the latitudes and longitudes corresponding to the maximum extent of the area where the data were collected. As noted above, more detailed geographical information should be provided (also) as a data entity in tabular or geographical format (e.g., GeoJSON).

temporal coverage Provide the earliest and most recent date for which data corresponding to the data submission were collected.

taxonomic coverage If relevant, list the Latin and/or common names of all organisms that are referenced in the data.

tab: personnel

List all personnel that should be given credit on the dataset. There are three role types: creator, associated_party, and metadata_provider. The creator type corresponds to dataset authors and will be included in the data package citation. The other types are supporting roles, and will be included in the metadata but not listed as data package authors. List creators in the order in which they should appear in the citation.

- creator: dataset authors; included in the dataset citation
- associated_party: indicates a person who contributed to the research but who is not a dataset author; in the metadata but not in the dataset citation.
- metadata_provider: indicates the person providing the dataset metadata, typically the person completing this form.

Though not required, we encourage strongly that all personnel, especially creators, should have and provide an ORCiD.

tab: 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 reuse the keywords found in related publications. CAP LTER data packages will by default be populated with a set of site-specific keywords, such as "lter" and "arizona". All keywords are ultimately rendered lower case.

For CAP LTER datasets only, at least one each `LTER Core Research Area` and `CAP LTER IRT` must be provided.

tab: data_entities

A dataset can contain one or many data entities, e.g., an individual tabular or geospatial data file. For tabular data, CSV file format is preferred but if you are submitting an Excel (or similar) workbook, each tab would be considered a separate data entity. Each data entity should be listed on its own row of the `data_entities` tab. Provide a description of each data entity.

It is good practice to not leave empty records. If the data contain missing values, document how these are coded in the `missing value code` column and provide a description of the missing value code in the `missing value code meaning` column. Typically, the same code is used to denote missing values throughout a data entity. For example, it would be rare to use NA and NULL in the same file. However, if more than one missing value code is used, document them in additional columns and be sure to mention this to the CAP LTER Information Manager.

tab: attributes

Here you annotate each attribute of each data entity in your dataset. The drop-down menu in the `data entity name` column is populated from the data entities you entered in the `data_entities` tab. For each data entity, list and provide a description of each field. If relevant, provide a unit of measure (e.g., meter squared) in the `unit` column. For non-tabular data, fields may not be an applicable concept but attributes may still apply (e.g., you would consider fields from the attribute table of vector data as attributes that should be documented here, similarly for the meaning of a cell in a raster data file).

tab: attribute codes

Codes apply to categorical variables or attributes that allow a specific set of values (R users will know these as factor levels). 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., do not use "M", "m", and "male" to refer to the same category).

tab: related_pubs

Here is a place to establish relationships between the data and any related publication(s), which is important for the audience of both data and papers. There are three possible relationships: (1) the dataset is citing the publication in some way, (2) if the publication uses the data, and (3) if the publication is *about* the data (i.e., a data paper).

If your publication(s) will come out after the data are archived, let the CAP LTER Information Manager know when the work is published as that connection can be made after both the data package and paper(s) are published.