Palmyra Atoll Data Library User Guide

Table of contents

Welcome!

The Palmyra Atoll Data Library (PADL) aims to publish as much data collected at Palmyra Atoll as possible.

We believe that facilitating access to research and its associated data will amplify academic discussions, help other researchers, and further inform management and policy decisions being made in Palmyra and other tropical ecosystems.

We are working with the Environmental Data Initiative (EDI), an NSF-supported initiative that provides a reliable, registered and certified trustworthy data repository with an open-access API. EDI follows a high metadata standard that ensures that your future self and others can utilize data for scientific and other forms of inquiry.

Why PADL?

We want to have a centralized place where scientist and managers have access to all data collected at Palmyra Atoll. We believe that data are among the most valuable outputs of research. If no-one can access these data, all the data collection is a shocking waste of resources. Having an open data library allows for aggregating and synthesizing data from different contexts. This is essential to establishing broader ecological knowledge and informing conservation management. Long-term data are crucial to understand historical patterns and baselines in a changing world. PADL wants to make it easy for scientist to know and have access to research done at Palmyra in the past.

Credits

This guide borrows heavily from EDI resources, CAP LTER Getting Started Guide and NCEAS coreR curriculum. Check out the links and see complete citation on the reference section.

About this guide

We created this guide to help you document your data in a way that facilitates the process of publishing your data or metadata to EDI. Here we describe what you need to do to make your data publicly available through the EDI repository.

❗ Important

We understand that EDI might not be suitable to all data. If you are planning to publish you data elsewhere we ask to please collect the metadata following EDI metadata standards presented in this guide and submit your metadata to TNC Palmyra.

What to expect

- The goal of this guide is for you to know what you need to do to be in compliance with TNC Palmyra Data Sharing Agreement.
- We provide some orientation and tips for when planning your data collection, together with resources to learn more about tidy data.
- We describe each of the elements needed to best document your data.
- This guide goes in-depth on how to publish your data using the ezEML tool suite developed by the Environmental Data Initiative (EDI).
- We also provide metadata templates for you to know and plan on documenting your data from the get go.

💡 Tip

We highly recommend to have a plan for collecting your metadata since the beginning of your research life cycle. According to EDI experience "continuous creation of metadata during the research life cycle greatly benefits data management during a project and when it comes time to publish data when the project concludes".

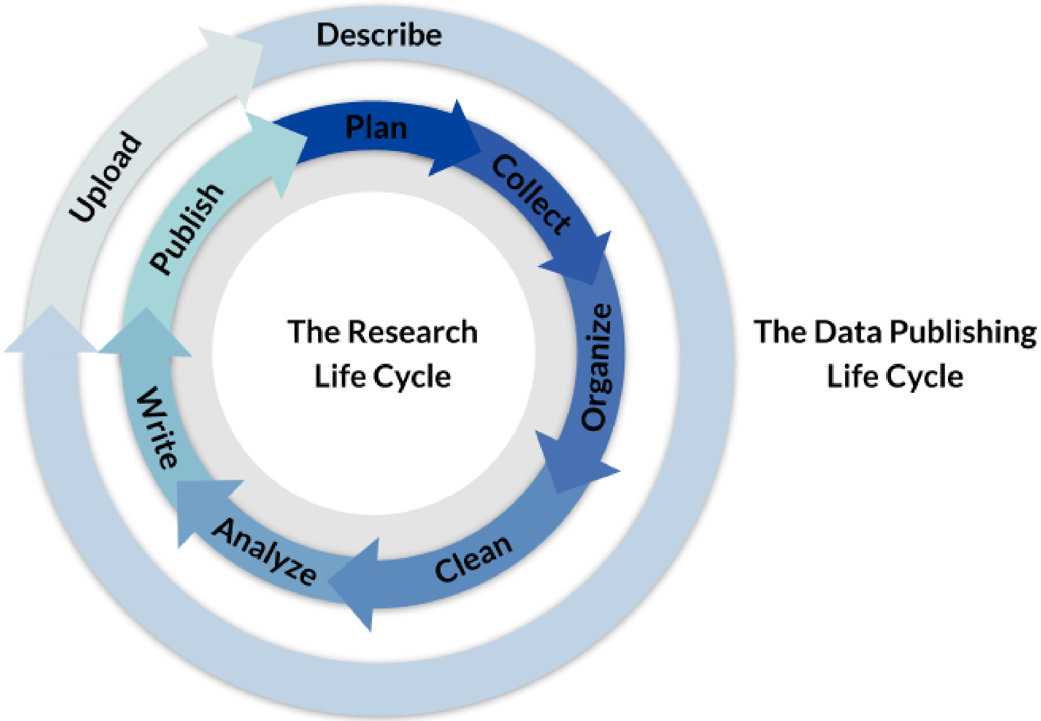

Figure 1: from edirepository.org, Creating Metadata During the Research Life Cycle

Basic Concepts

Before we start we want to introduce to some relevant concepts related to data publishing. This sets the ground to navigate the environmental/ecological data publishing world.

Data Package

- EDI publishes data from the ecological and environmental sciences irrespective of funding origin.
- The *unit of publication* is called a data package.
- This is an assemblage of science metadata, the data it self, and any other file you want to publish with your data.
- The basic rule of thumb is to package your data in the form you would like to receive it.
- Each data package is assigned a Digital Object Identifier (DOI) and published in the repository for future use.
- EDI resources on data packages

Metadata

- The metadata is the data that provides information about your data. It describe the structure, content, and context of your data.
- They are vital to the discovery and reuse of data, and are a required element of a data package.
- EDI resources on Creating Metadata for a Data Package

Ecological Metadata Language (EML)

- EDI uses the Ecological Metadata Language (EML) format to document the metadata in each data package.
- "EML provides a comprehensive vocabulary and a readable XML markup syntax for documenting research data. It is a community-maintained specification, and evolves to meet the data documentation needs of researchers who want to openly document, preserve, and share data and outputs" (Jones et al, 2019).

ezEML

- EDI created a user-friendly tool that streamlined the process of the creating metadata in the Ecological Metadata Language (EML).
- ezEML is a form-based online application that leads the user through EML document creation step by step.
- Check out EDI ezEML User Guide for more information.

OCID ID

"a free, unique, persistent identifier for researchers. Having an ORCID distinguishes researchers with similar names, ensures that scientific output is correctly attributed, and improves the overall discoverability of scientific products. Linking data publications to an ORCID profile ensures that research contributions are correctly attributed across scholarly databases."

- We highly recommend creating a ORCID
- You can use your ORCID t log in to ezEML **FROM EDI NEEDS TO BE REVISED**

Data Workflow

1. Organize
2. Clean
3. Describe
4. Upload
5. Cite

Metadata Components

This section describes of the most relevant components of the EML. The goal is for you to learn about all information you need to document from your data from the get go. The next sections show formats that can help you do this along the way.

Title

Every data package needs a **Title**. Data packages title should be informative and provide helpful context about the dataset. Title should inform about the temporal and geographic extent of the data.

> ℹ Example
>
> Activity budgets and space use for two common Pacific parrotfish (Chlorurus Sp.) at Palmyra Atoll National Wildlife Refuge 2014.

Data Tables

This are all the tabular data to be included in your data package, saved as a CSV (comma-separated value) text file. Each data table in your package needs to be documented and described with the the following information.

1. General information of a data table

- Name of the data table
- Short description
- File name

Following our example, the table below shows the information for the three data table in the parrotfish data package.

Example

Data table title	Data table description	File name
Chlorurus Activity Palmyra 2014	Reports the start and stop time of each focal follow as well as the start and stop time of each activity for C. Microrhinos and C. spilurus	Chlorurus_Activity_Palmyra_2014.csv
Fish Information Chlorurus Data	Total length and phase for each individual in our study	Fish_Information_Chlorurus_Data_20

2. Column properties It is necessary to describe the content of each data table with the following information.

- Column name
- Definition of the column (what is each variable)
- Type of data (Numeric, Text, Categorical, DateTime)
- If numeric provide units
- The precision of measurement (optional)
- Format of dates (eg. YYY-MM-DD)
- If there is missing values how are they coded (optional)
- Explanation of why are there missing values (optional)

Tip

ezEML allows you to upload your data file to assist in generating the metadata describing each table. For example, it identifies each column name and the *type* of data in each column.

3. Defining codes in categorical variable Each of the unique categories or codes in a categorical variable needs a description about what they mean.

Participants and Personnel

There are several personnel roles associated with a dataset, including Creators, Contacts, Associated Parties, and Metadata Providers. The form for each role is basically the same, and allows providing the usual details, such as name, salutation, and address.

The categories for these sections are: - Last Name - First Name - Middle Name - ORCID ID - Organization - Organization id - Position Name - Email - Address - Phone

You can provide information as desired but please be sure to at least provide full name, email, organization and ORICI ID if available for each person.

Creators

Creators are the dataset authors and/or primary project contributors. Dataset creator is equivalent to the author of a journal article or book. Just as with a journal or book, order matters and the creators should be listed in the order to which they contributed to the project or dataset. These name will be part of the citation of the data package.

Contacts

Any correspondence about the dataset will be sent to this person.

Associated Parties

People who contributed to the project in some way but who are not considered dataset authors. Someone who helped with some of the field or lab work, or who were involved with the overall project but not this particular dataset would be good examples of an associated party.

Metadata Providers

You! Presuming that you are a dataset author. Or anyone that is well versed witht the data to provide all the detailed metadata.

Abstract

Description and overview of the data package. This abstract is analogous to the abstract of a journal article, but specifically the data. There are not character limitations so please be generous with details. Include what, why, where, when, and how of this data package.

> ⚠ Warning
>
> You cannot use the same text from the abstract of a published work as the abstract of your data package. This can cause copyright infringement. If your data package is associated with a journal article, your data package abstract need to be different than the one in the article.

ℹ Example

Data was collected at Palmyra Atoll National Wildlife Refuge located in the Central Pacific in 2014, and the work focuses on two species of Parrotfish, Chlorurus microrhinos and Chlorurus spilurus (formerly Cholrurus sordidus). For C. microrhinos, the project was designed to collect fine-scale spatial behavior data, focusing on territory size, species interactions, and benthic impact (i.e. feeding behavior). Focal follow data was collected by one or two observer(s), either on snorkel or SCUBA, and recording focal activity down to the second. Simultaneously, the observer would be towing a GPS that was recording a location every 5 seconds and each location was then associated with a particular behavior. Throughout this study individual fish were identifiable and successive follows were possible on individuals. For C. spilurus the focus of the project was to collect behavioral time budget data on feeding, territorial defense, and spawning behavior. The 'Chlorurus_Area_Palmyra_2014.csv' data only covers C. Microrhinos and gives the 95% KUD area estimation for GPS towed tracks, as well as the 95% KUD area for locations where feeding was occurring. We also report the step length between successive points for the entire follow as well as where feeding was occurring.
The 'Chlorurus_Activity_Palmyra_2014.csv' is for both C. Microrhinos and C. spilurus and reports the start and stop time of each focal follow as well as the start and stop time of each activity. Activity descriptions can be found in the metadata and the total length and phase for each individual in our study can be found in 'Fish_Information_Chlorurus_Data_2014.csv.

Keywords

You can add as many keywords as necessary associated to your data package. Using keywords from a controlled vocabulary (CV) will improve your data's future discovery and reuse.

The LTER CV is a good source for keywords. Access the LTER CV here. Also, please determine one or two keywords that best describe your lab, station, and/or project (e.g., Trout Lake Station, NTL LTER).)

PADL request to please include **Palmyra Atoll** as a keyword so we can keep track of all the data package in our library.

Intellectual Rights

EDI requires for your to choose under which licence you want your data to be published. This are both Creative Common licenses which allow open sahring of your data

The options are:

- Creative Commons CC0 1.0 “No Rights Reserved” - Less restrictive, no copy rights. You dedicate your data to the public domain.
- Creative Commons license - Attribution - CCBY - The data package is open to be shared and adapt. Anyone who use this package must give appropriate credit.

Geographic Coverage

Here you present the physical location associated either with data collection (e.g., field work), data analyses (e.g., spatial extent of a model), or both. You can provide this information in different ways.

1. Provide a single geographic description and set of bounding coordinates. In the case of the Palmyra Atoll, your would provide a single set of bounding coordinates that would encompass all of the sampling locations or complete Atoll.
2. Provide separate, distinct geographic locations for each sampling location. Here you can provide coordinate of a specific point and describe which isllet on the Atoll the sample is from. You would repeat this process for each sampling location.

The second option is more information rich but not always necessary or relevant. Either approach is acceptable.

Coordinates must be provided as latitude and longitude in decimal degrees.

Temporal Coverage

Information about when the project stared and when it ended. This should be the dates associated only to the data collection in this data package, , and should not include time spent, for example, analyzing data. You can provide a specific day or a year.

Taxonimic Coverage

“If relevant (do not consider humans), provide details of the taxonomy of organisms featured in the study. You can provide these metadata through the Taxonomic Coverage link in the Contents menu.

For each taxon, highlight the scientific name and taxonomic resolution of the organism using the Taxon Scientific Name and Taxon Rank drop-down lists, respectively. Details of most organisms are available through ITIS but draw upon different taxonomic authorities with the Taxonomic Authority drop-down if appropriate. Once you have identified the details, use the

Fill Heirarchy button to have ezEML construct a full taxonomy for the organism. Save and Continue, then repeat for each taxon."

FROM CAP LTER! NEED TO ADAPT

Maintenance

Mention here if the data collection is completed or ongoing. If it is an ongoing project, it is recommended to mention how frequently this package is going to be maintained. If needed, you can add any other necessary information about the maintenance of the data package.

Methods

Describe all methods used to collect, process and analyze the data. It is crucial to provide detailed information for potential users to be able to interpret the data correctly for reuse. Be specific about the study design and field and lab methods for collecting and processing the data. Include instrument descriptions and protocol citations.

Example

Focal follow data was collected at two locations at Palmyra Atoll National Wildlife Refuge, Penguin Spit and Western Terrace. Individual fish were identified by unique markings on the face, tail, and body as well as missing scales or scars. A diver would identify an individual fish and follow the fish for 2-3mins before stating the focal follow. The diver would tow a GPS that was recording a position every 5 secs and synchronized the time on their wrist watch to the GPS time. The diver would then follow the fish and record the start and stop time of each activity, while also following the path of the fish. GPS tracks were downloaded and each position was associated with an activity and activity summaries were calculated. For C. Microrhinos we calculated the 95% kernel utilization distributions (KUD) for the entire focal follow as well as for only the locations categorized as 'feeding'. All KUD estimates were done in R with the adehabitatHR package.

Project

If desired here you document the information about the project this data is part of. Filed in this section are

- Project title

- Project abstract
- Funder name
- Award title
- Funder ID
- Award url

You can also add detailed information about project personnel. Fields in this section are similar to the Participants section.

Other Entities

Any other file that is not a data table (tabular data) can be part of your data package as “other entity”. This may be such things as images, zip files, R or Python scripts or any number of other items that are not tabular data files.

⚠ Warning

MENTION SOMTHING ABOUT SPATIAL DATA

For each entity you want to include in your data package you have to define, - The type of file (generally the extension of the file, eg and R script would be R) - Short description about the entity - Name of the file - Format

Tip

Similar to Data Tables, your can upload your file to ezEML and it will aromatically identify some of the metadata for this entity.

Using ezEML

Metadata Tamplates

Alternative options

Additional Resources