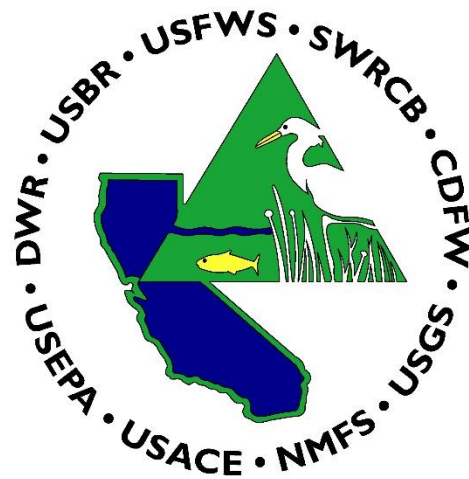


# Step-by-Step Guide to Creating Descriptive Metadata (EML) with ezEML

*Version 1.0*



## Interagency Ecological Program

COOPERATIVE ECOLOGICAL  
INVESTIGATIONS SINCE 1970

*Last Updated: October 10, 2024*

## ***Suggested citation***

Data Publication Working Group. “Step-by-Step Guide to Creating Descriptive Metadata (EML) with ezEML”. Version 1.0. Interagency Ecological Program, 10 Oct. 2024, <https://iep.ca.gov/Data/Data-Utilization-Working-Group>. Date of Access.

Version Number	Date Created	Description of changes	Justification for changes	Version editor(s)	Contact information
1.0	09/18/2024	Document created		DPWG	trinh.nguyen@wildlife.ca.gov

## **Acknowledgements**

Thank you to Rosemary Hartman, Sam Bashevkin, Celeste Dodge, Vanessa Mora, and Morgan Gilbert for their helpful comments in making this guide more usable and useful.

# TABLE OF CONTENTS

Introduction .....	4
Workflow .....	5
Preparation.....	5
Logging into ezEML.....	5
Starting an EML document .....	5
Essential features of ezEML.....	6
Filling out ezEML to create your EML document .....	7
Title.....	7
Data Tables .....	7
Creators .....	9
Contacts.....	9
Associated parties.....	9
Metadata providers .....	9
Abstract .....	9
Keywords .....	9
Intellectual rights .....	10
Geographic Coverage.....	10
Temporal coverage .....	11
Taxonomic Coverage.....	11
Maintenance.....	12
Methods .....	13
Project .....	14
Other Entities.....	14
Data Package ID .....	14
Final Steps.....	14
Finalizing your package .....	14
‘Submit Package to EDI’ .....	14
‘Collaborate with Colleagues’ .....	15
Exporting EML data.....	15

# Introduction

**Purpose:** This guide walks users step-by-step through the ezEML tool to create detailed metadata. Metadata provides critical information about your data package that allows users to more easily discover, access, and interpret your data.

**What is it:** ezEML is a web tool created by the [Environmental Data Initiative](#) (EDI) to simplify the creation of Ecological Metadata Language (EML) metadata. The tool guides you through a series of fillable forms, prompting you to enter specific metadata in each. As you fill out your information, built-in automatic checks will alert you to potential errors to address.

Once all metadata fields are completed, ezEML provides a direct connection for uploading your data package to the EDI data curation team. The team performs final checks prior to publication to the EDI Repository. Alternatively, you can download the formatted EML metadata separately to publish to a different data repository.

**Scope:** This guide assumes that data quality assurance and quality control (QAQC) have already been completed. It focuses only on the metadata creation process immediately before data publication.

**Workflow:** We define the data publication process as three main steps. This guide informs the second step:

1. QAQC and prepare your collected data
2. Fill out all fields outlined in ezEML (this guide informs this step)
3. Publish to EDI

**Usage:** The major sections of this guide mirror those of the ezEML application. We recommend that users follow this guide sequentially as they are working in ezEML, carefully considering all additional details in each section.

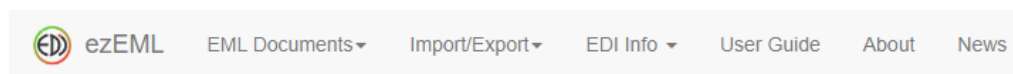
# Workflow

## Preparation

We recommend you dedicate a folder to your publication workflow. You will download and upload files to and from this folder throughout the publication process. We also recommend that your data files be in flat-file formats, e.g., CSV, as these are non-proprietary, widely used, and have a dedicated section in ezEML. However, you can still upload files of other formats onto EDI, and we recommend you do so if they are valuable, in the “Other Entities” section of ezEML.

## Logging into ezEML

1. You can access the website at <https://ezeml.edirepository.org/eml/auth/login>
2. There are many ways to log in, e.g., a Google account, and you can use any. You do not specifically need an EDI account.
  - a. If you do want an EDI account, you must contact EDI ([support@edirepository.org](mailto:support@edirepository.org)). The main reason to create one is if you would like to automate the publication process via ‘EMLAssemblyline’ (an R package, <https://github.com/EDIdorg/EMLassembleline>).
3. Once logged in, the website contains several important tabs at the top left. A screenshot of these tabs is shown below and described in subsequent bullets:



- a. EML Documents
  - i. Allows you to create, open, or manage EML documents created via ezEML
- b. Import/Export
  - i. To open or save completed EML documents
- c. EDI Info
  - i. Knowledge resources about the repository and its services
- d. User Guide
  - i. Relevant guides on how to use ezEML. We have attempted to incorporate all necessary information in this current guide; however, users are encouraged to explore the additional guides in this tab for additional information.

## Starting an EML document

1. From the ‘EML Documents’ tab, create a ‘New’ document or ‘Open’ an existing document
2. Existing documents
  - a. If you are updating a data package you have published previously, you can open the existing file to make updates.
  - b. If you are a collaborator, you can find your document under the ‘Collaborate’ tab in the top right corner. Collaborators are capable of editing information within ezEML


- just as the main creator. In the past, if you ask EDI to publish your dataset for you, i.e., you provide them with a metadata word template, EDI will be the main creator and will list you as a collaborator.
3. Follow and fill out the outlined sections of ezEML, ideally in parallel as you are working through ezEML. We have mirrored the sections below with the order of sections in ezEML.

## Essential features of ezEML

### Contents

-  Title
-  Data Tables
-  Creators
-  Contacts
-  Associated Parties
-  Metadata Providers
-  Abstract
-  Keywords
-  Intellectual Rights
-  Geographic Coverage
-  Temporal Coverage
-  Taxonomic Coverage
-  Maintenance
-  Publisher
-  Publication Info
-  Methods
-  Project
-  Other Entities
-  Data Package ID

Table of Contents of sections within the ezEML tool. Each section is hyperlinked to provide you to easily navigate between sections. A colored, circle marker for relevant sections indicates if there are errors (red), warnings (orange), or no errors (green). Clicking on a marker will provide additional details on the existing warnings and/or errors.

1. Contents menu: this is a Table of Contents that outlines all fillable metadata sections of ezEML. You can navigate from one to the next using the 'Save and Continue' button or by clicking the section name directly.
2. Colored indicator: certain sections have colored indicators next to them. These are from error checkers that run in the background as you fill out your metadata. A red icon indicates errors exist, an orange indicates warnings to be aware of, and a green icon indicates no errors. All red icons must be addressed before publication. Clicking on the icon will provide a list of errors and warnings after which you can click a specific item that takes you directly there. Certain sections will not have an icon; this does not mean they are unimportant but just that there are no automatic checkers.
  - a. These colored indicators also show up throughout other pages of ezEML.
3. Help bubbles: many sections and fields have a help button that provides valuable clarifications about a field. 

## Filling out ezEML to create your EML document

### Title

1. The title of your data package. If your package is an IEP survey, the title should begin with “Interagency Ecological Program:” and followed by the name of your sampling program (as recommended by [IEP Metadata template](#)).
2. Ideally, use descriptive title that includes the sampling program (who), location (where), time frame (when), and types of variables included (what). For example: “Interagency Ecological Program: Fish catch and associated water quality for the Fall Midwater Trawl, Sacramento-San Joaquin Delta, 1970-2023”.

### Data Tables

1. Upload your data via the ‘Load Data Table from CSV File’ button. We **do not** recommend adding tables from scratch.
2. This section supports only flat-file data files, e.g., CSV files. If you have other file types, you can upload and describe them in the “Other Entities” section (discussed later).
3. Add a short description of the data table after upload that easily identifies what it is.
4. Edit column properties
  - a. This step allows you to provide detailed metadata about your data tables.
  - b. Ensure that the ‘Type’ (data type) for your columns is correct (ezEML tries an initial guess). If incorrect, change it via ‘Change Type’. It is important that this is the first thing you do during this editing step. There are four types of data types:
    - i. Categorical: values come from a fixed set of defined categories. Each unique category will need to be defined.
    - ii. DateTime
    - iii. Numerical
    - iv. Text: Values can be any text (e.g., unconstrained notes)
  - c. We recommend using the ‘Download Column Properties Spread’ feature to fill out your data table metadata.
    - i. This downloads a formatted Excel file. Save it to a folder in your dedicated publication folder. You will upload this file to ezEML after filling it out.
    - ii. This Excel sheet allows you to fill in metadata for all columns across your data tables at once.
      1. This is more efficient than editing each column properties in the online tool, although you are free to do it online too.
    - iii. The colored cells in the sheet are recommended or required cells: they are
      1. Missing value code:
        - a. Fill this out only if you have a value dedicated to representing missing data, otherwise, leave blank
        - b. ezEML may fill out rows for this column if it detects certain strings in your column, e.g., “9999” or “NA”. Make sure to

delete if this is not the case, e.g., an AutoNumber column containing 9999.

- c. For any specified value, you must also add an explanation in the neighboring cell.
2. All 'Numeric' columns require a unit label
  - a. Each 'Numeric' column is a row on the Excel sheet, and you should scroll up and down to find them all.
  - b. A 'Standard Unit' can be chosen from a drop-down menu
  - c. For all other units, enter the unit in the 'Custom Unit' cell
    - i. At times, units of a different magnitude may not be fully represented (e.g., centi- is included but not micro-), so you will specify it as a custom unit.
  - d. For all columns with no units, indicate "dimensionless" as a custom unit. This is due to a blank entry being ambiguous – does it mean no units or simply forgotten?
3. All 'Categorical' columns need their categories defined
  - a. Each 'Categorical' column is a n x 2 (row x column) table on the Excel sheet, and you should scroll left to right to find them all.
  - b. ezEML will try to pull out your categories for you to define. Ensure that all are represented (it may miss categories that theoretically exist but were not actually logged).
4. All 'DateTime' columns need their format defined
  - a. Each date-time number is represented by a letter separated by a symbol (case sensitive):
    - i. "Y" = year
    - ii. "M" = month
    - iii. "D" = day
    - iv. "h" = hour
    - v. "m" = minute
    - vi. "s" = second
  - b. For example, "2024-01-01 12:00:00" is represented as "YYYY-MM-DD hh:mm:ss". Note the space between date and time.
  - c. EDI prefers date time to be formatted as "YYYY-MM-DD hh:mm:ss" or a subset of it (e.g. "YYYY-MM-DD" for dates only)
- iv. Once filled out, use 'Upload Column Properties Spreadsheet'
  1. ezEML will parse your metadata from your file after the upload
  2. All columns should theoretically turn green. If not, click on the icon to understand why and navigate to the affected item(s).



## Creators

1. The people and/or organizations responsible for creating the resources. For IEP datasets, please list “Interagency Ecological Program” as a creator as an “Organization”.
2. Order is significant and indicates contribution.

## Contacts

1. The people or organizations to be contacted for additional information. This generally includes at least one creator.

## Associated parties

1. The people or organizations involved in any way collecting and/or creating the data or data package (e.g., technicians, agency collaborators)
2. Optional. These individuals will be displayed below the ‘Creators’ and ‘Contacts’ fields

## Metadata providers

1. The people or organizations involved in creating the metadata content.
2. Optional. You would fill out this section only if there are specific details that only the metadata provider(s) know, if they are not the same as the creator(s)
3. The EDI curation team does not need to be cited

## Abstract

1. The abstract should contain enough information to allow your users to understand critical aspects of your data package, such as:
  - a. What question(s) is your data package meant to address?
    - i. Why is this dataset important?
  - b. How is the data collected, its methods, e.g., midwater trawl?
  - c. What is the geographic extent of the dataset, e.g., Suisun Marsh?
  - d. What is the temporal range of the dataset, e.g., date range and sampling season?
  - e. What type of data is in the dataset, e.g., fish count and water quality?
    - i. Include what species, if applicable
    - ii. Advertise key variables
  - f. What format can the users expect, e.g., flat files such as CSVs?
  - g. How can the dataset answer the questions the data package is meant to address?

## Keywords

1. You need a minimum of 5 words to pass the checker but adding more could make your dataset more findable.
2. If you are publishing an IEP dataset, include:
  - a. “IEP”
  - b. “Interagency Ecological Program”
3. We recommend you include abbreviations associated with your dataset, e.g., FMWT, EMP, or DJFMP.

4. Controlled vocabulary
  - a. A controlled vocabulary contains a standardized list of keywords relevant to a topic.
  - b. Using keywords from a vocabulary may allow users to more easily find your dataset, i.e., words commonly used by a field or community.
  - c. ezEML offer the Long Term Ecological Research Network (LTER) keyword list as a controlled vocabulary. We recommend that you use keywords from this list if possible.

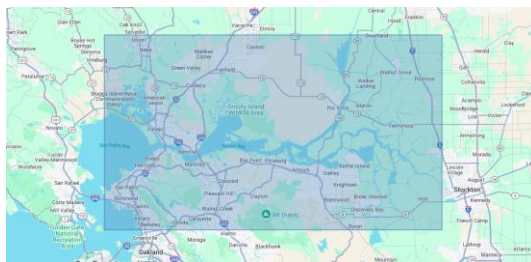
## Intellectual rights

1. This section explicitly tells users if they need to cite your data package for use or not.
2. ezEML provides two default choices:
  - a. Creative Commons CC0 1.0: “No Rights Reserved”
    - i. Users are free to use the data as they see fit
    - ii. Citation is recommended but not explicitly required for use
  - b. Creative Commons License – Attribution – CC BY
    - i. Users can use the data as they see fit but must cite the package
3. We recommend you use the latter, ‘Creative Commons License – Attribution – CC BY’. This license provides users with full flexibility when using the data while providing credit to the creators.
  - a. Note that only individual(s) and/or organization(s) listed in the “Creator” section will be listed as authors when users cite the package.
4. You are free to use other licenses if they better suit the requirements of your organization.

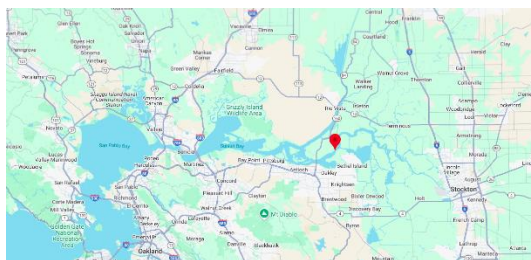
## Geographic Coverage

1. Define the total spatial extent of your data package: use ‘Add Geographic Coverage’
2. ‘Geographic Description’: list out the major regions in your study, e.g., “San Pablo Bay, Suisun Bay, and Suisun Marsh”
3. West, East, North, and South bounding coordinates: enter coordinates to visualize the spatial extent of your study. West and east coordinates are your minimum and maximum longitudes, and north and south coordinates are your minimum and maximum latitudes, respectively. If you publish to EDI, this is shown as an interactive map. There are two ways EDI can visualize your geographic coverage:
  - a. Bounding box: a box that encompasses all your sampling locations. Use this option if you have many sampling locations across a large geographic area.
    - i. To display this, enter the four coordinates associated with the corners of your bounding box. You can use various tools to get these coordinates, e.g., ArcGIS or `st_bbox()` from the ‘sf’ R package.
      1. This is a simple [webtool](#) that lets you specify your own bounding area. Select “DublinCore” as the output (bottom left) to receive the coordinates in decimal degree format for use in ezEML.

- ii. For example, a bounding box for the 20 mm is recorded as: -122.4149, -121.3686, 38.3335, 37.859 (W, E, N, S). You can view your bounding box on the app via 'Preview', as shown below:



- b. Individual points: markers to represent your individual sampling locations. Use this option if you do not have many sampling locations or would like to emphasize exactly where they are.
  - i. To display this, enter the same longitude coordinate for the W and E options and the same latitude coordinates for the N and S options. Essentially, a bounding box is 4 distinct values while an individual point is only 2 distinct values.
  - ii. For example, a point is recorded as: -121.693, -121.693, 38.05378, 38.05378 [(W, E, N, S), note the repeated values equating to only 2 distinct values)]. You can view markers on the app via 'Preview', as shown below:



- 4. You can import geographic coverage from another package ('Import Geographic Coverage') or load your points from a CSV file ('Load Geographic Coverage from CSV file'). Refer to the help balloon in ezEML for additional details.

### Temporal coverage

1. The first and last sampling dates of your data package.
2. The ending date for an ongoing survey can be left blank or with the last date of sampling.

### Taxonomic Coverage

1. This information allows users to find your dataset when searching for specific species of interest.
2. We recommend you use 'Load Taxonomic Coverage from CSV File' to load all species at once

- a. You must create the CSV file yourself. It must contain three columns named exactly:
  - i. `taxon_scientific_name`
    1. These must be spelled correctly. ezEML provides three taxonomic authorities to choose from:
      - a. NCBI – [National Center for Biotechnology Information](#)
      - b. ITIS – [Integrated Taxonomic Information System](#)
      - c. WORMS – [World Register of Marine Species](#)
    2. We recommend using the NCBI database
  - ii. `general_taxonomic_coverage`
    1. General comments about the taxon. Usually left blank
  - iii. `taxon_rank`
    1. A specific taxonomic level to be displayed for a species. Usually left blank
  - iv. For example:

	A	B	C
1	<code>taxon_scientific_name</code>	<code>general_taxonomic_coverage</code>	<code>taxon_rank</code>
2	Catostomus occidentalis		
3	Tridentiger bifasciatus		
4	Ameiurus catus		
5	Ictalurus punctatus		

1. Most databases already include a table with the species code and scientific names. You can aggregate to unique names to copy into this sheet.
- b. Upload the CSV file and ezEML will pull and format the relevant taxonomic information from the chosen taxonomic authority database.
  - i. This process can take a long time if you have many species or if your internet speed is slow.
  - ii. The process may return an error page after completion. Try reopening the in-progress data package again in ezEML.

## Maintenance

1. We recommend you include the following information in the 'Description' box:
  - a. Name(s) of the people maintaining the data package, generally one of the creators
    - i. Contact information
    - ii. Link to the program webpage, if applicable
  - b. All locations that the data can be found, if applicable
2. Maintenance Update Frequency: we recommend that you choose an option from the drop down despite this being optional. This allows users to set their expectations.

## Methods

1. We recommend that you dedicate a method step in ezEML (created from 'Add Method Step') to each major section in your methods (detailed in bullet 2 below). You should approach writing the methods as you would for a journal article.
  - a. Alternatively, you can upload a supplemental PDF describing your methods, however, this option is less preferred due to being less formatted and machine-readable. Provide links to SOPs or other references as needed.
2. Method steps commonly include:
  - a. Study design and objective
    - i. What is/are the ecological question(s) of interest? Why was this research undertaken?
  - b. Study area
    - i. Describe key locations
    - ii. What are the general habitats sampled, e.g., wetlands, shoals, channels?
  - c. Data collection
    - i. Sampling design, e.g., fixed or random stations, midwater trawl, oblique tows
    - ii. Sampling season, e.g., September through December
    - iii. Sampling gear, e.g., Kodiak net, 20 mm net, 500-micron net
    - iv. Variables collected, biotic and abiotic
    - v. Identify established protocols if existing
  - d. Quality assurance and quality control (QAQC)
    - i. Any procedures associated with the QAQC process, before and after data entry, e.g., genetic testing, double entry, line-by-line
  - e. Data processing and management
    - i. Data entry protocols
      1. When is data entered, e.g., electronically, after sample, after survey, after season?
      2. Method used, e.g., electronic, manual, double entry
      3. Data storage protocols, e.g., a program (front end) to update the database (back end), update the back end directly, etc.
    - ii. Data processing protocols
      1. Any steps after collecting the sample to prepare for data collection
      2. Any manipulations of the stored data in preparation for publication
    - iii. Project history
      1. Major changes in methods and sampling locations with dates of these changes
  - f. Data publication
    - i. What format is the attached data in, e.g., relational tables, integrated table?
    - ii. File format, e.g., flat files like a CSV, RData

## Project

1. Project Title: a project documents the research context of your data package. Use a descriptive title. This is different from the “Title” section which describes the “what”, “where”, and “when” of your data package; the “Project” section details the “why” and “how” of your data package, e.g., Species Diversity of Fishes in the San Francisco Bay-Delta Estuary.
2. Project personnel: optional. Usually covered in the ‘Creator’, ‘Contacts’, and ‘Associated Parties’ sections.

## Other Entities

1. An entity is any other files uploaded that are not data tables (which got their own section). These additional files require metadata related to their name, type (e.g., photograph), filename, and format. We recommend you provide a short description of the file in the ‘Description’ field.
2. Use ‘Load Other Entity from File’ to upload your additional files. You can then ‘Edit’ the metadata associated for each of them on the webpage.
3. These additional files can be any file you feel will be useful to your users, e.g., an SOP, a log of changes, associated publications, an Access database, etc.

## Data Package ID

1. A unique ID assigned to your data package for identification purposes.
2. If you are publishing to EDI, EDI will fill this out for you.
  - a. You can fill this out yourself but must reserve a Data Package ID before doing so. EDI provides this service and will provide you with an unused ID, e.g., edi.1118.1. Generally, you would only do this if you wanted to automate your publication process through the EMLAssemblyline R Package.

## Final Steps

1. Click on ‘Check Metadata’ to ensure that no sections are red and are, preferably, green. You can click on specific links to go directly to the affect field. Fix all errors (red). If there are warnings, make sure that you are ok with them. Warnings will not prevent publication, but errors will.
2. Ensure that ‘Check Data Tables’ is not red. The same process described in ‘1.’ above applies here.
3. ‘Explore Data Tables’: this is an optional tool that provides you with autogenerated information about your data tables.

## Finalizing your package

### ‘Submit Package to EDI’

1. This sends the package to the EDI Curation Team. This step does not constitute publishing your dataset--it is a step prior.

- a. This team will proof your package for critical errors (if they still exist) in preparation for publication. Although this is a very helpful service to have, you should not rely on this step to catch errors.
  - b. If you are submitting an update to a publication, be sure to mention the package ID in the comments section.
2. The curation team will publish your data package to a 'staging' (draft) server and get your feedback. EDI will only publish your dataset to the public facing server with your direct confirmation.
  - a. The 'staging' server works exactly like the public server but simply embeds the webpage with "Test Data Package" watermarks to indicate a pre-publication status. This staging website will display your data package as it will display on the publication site.
    - i. This is the best time to ask others, e.g., your co-authors, the DUWG, or the DPWG, to review your package.
    - ii. EDI may ask for changes if they catch any issues in your package.
  - b. If you have changes you would like to make after seeing your package on the staging server, make them from the same ezEML file. Resubmit the package once done.
  - c. EDI generally respond within 1-3 days if not earlier. If you do not get a response within 5 days, send a follow-up email.

### 'Collaborate with Colleagues'

1. This invites others to make direct changes to your data package within the ezEML file. We **do not** recommend this option. The main drawback is that only one person can work on the package at once and changes may not be communicated well between collaborators.
2. EDI will be listed as a collaborating colleague once you submit your ezEML. If the curation team is working on your package, you may not be able to access it. Generally, you should simply wait for them to finish, otherwise, email them.

### Exporting EML data

1. This option is an advanced use and is generally not applicable. You can export the data you have filled in as an XML file. This is done via the 'Import/Export' tab. There are two options:
  - a. Your entire zipped package. This will include the XML file containing your metadata and any supporting files, e.g., data files, that you have included.
  - b. Only the XML file
2. This option is useful if you need formatted EML metadata to publish onto another repository, e.g., KNB Repository.

### Fin

Congratulations on reaching this milestone! Your meticulous work documenting detailed metadata about your data package is an invaluable contribution to understanding and preserving the natural resources around us.

Specifically, by making your data package readily discoverable and accessible, you are maximizing its potential to inform critical management decisions. Your efforts may contribute to the recovery of endangered species, the protection of essential habitats, or the development of more effective management strategies.

Well-documented data also becomes a vital building block for future research and helps foster a collaborative environment where scientific advancements build upon one another. In such an environment, your thorough metadata documentation ensures the long-term integrity and trustworthiness of your dataset.

Additionally, your commitment to comprehensive metadata creation is a significant investment in saving time and resources for both yourself and data users in the future. You, as the most knowledgeable individual to do so, are uniquely positioned to ensure the utility of the data for years to come.

And finally, you are setting a powerful example of data stewardship. You are helping to promote a cultural shift within our community to recognize the essential role of data sharing and detailed documentation in advancing scientific understanding and effective management of our environment. Thank you for your dedication to data stewardship and your contributions towards safeguarding our interconnected natural and human systems.