

Introduction to Data Management and Publication

By Gabriel Kamener
FCE Information Manager, Florida International University



2025 Florida Coastal Everglades Information Management

INFORMATION MANAGEMENT
Feb 18, 2025

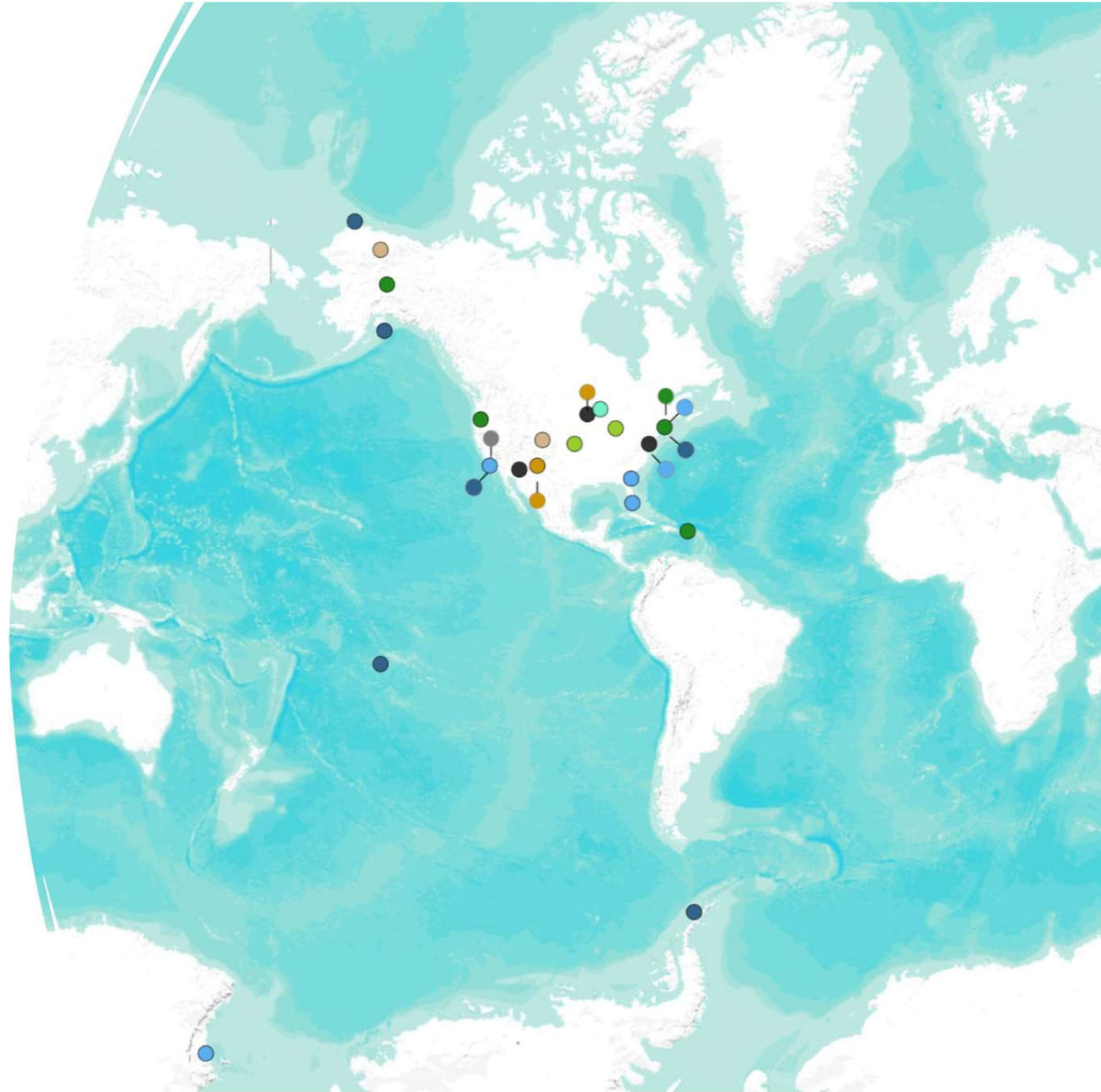


Overview

- Review
 - Sharing of LTER Network data
 - EDI data repository
- Best practices
 - Data and project management
 - Formatting data
 - Describing metadata
- Preparing for publication with ezEML

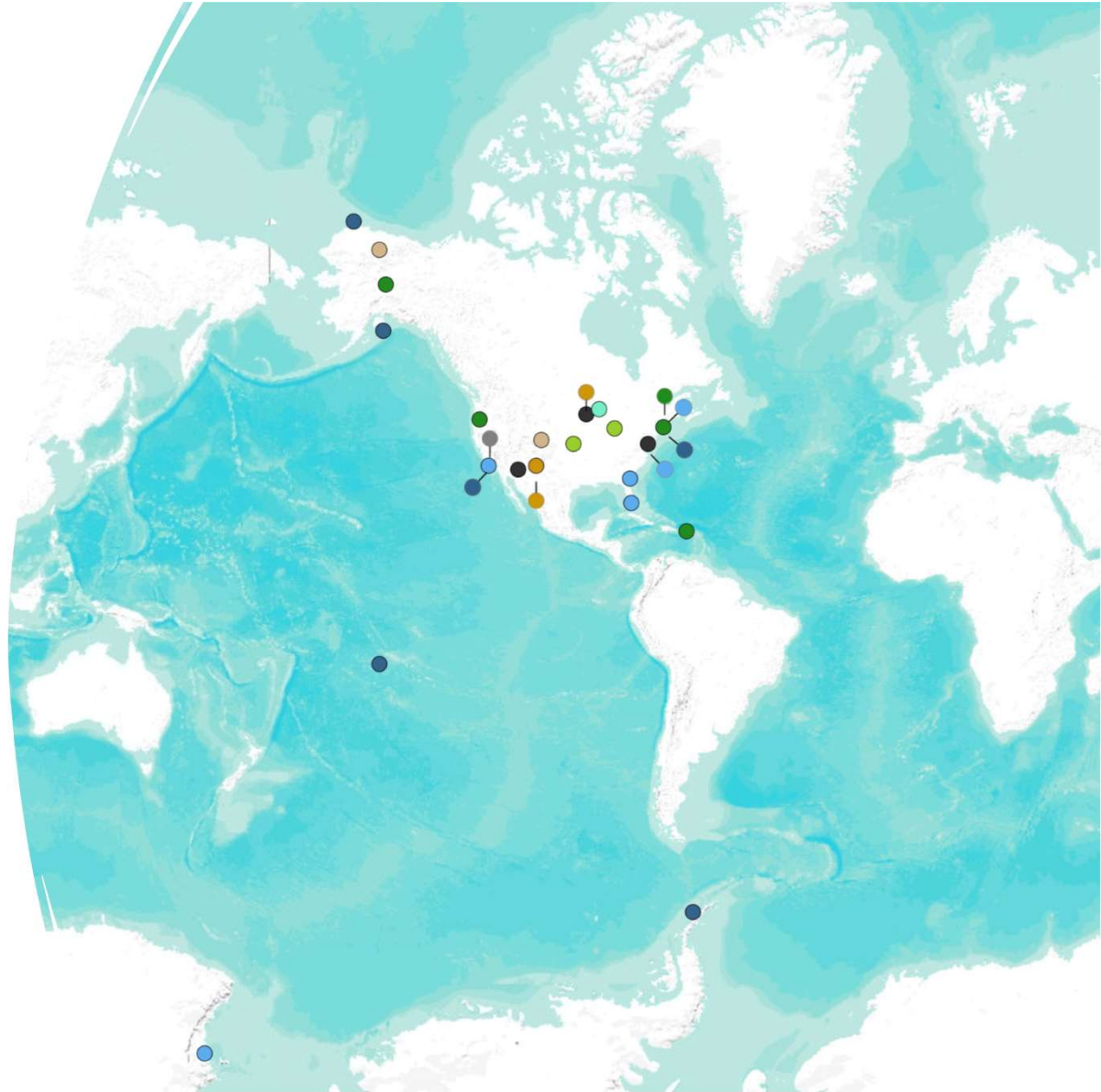
Sharing of LTER Network Data

- Ongoing since NSF funded first LTER sites in 1980
- Enables new science!
- Supports open science and reproducibility
- Funders and journal publishers often require datasets in a recognized data repository



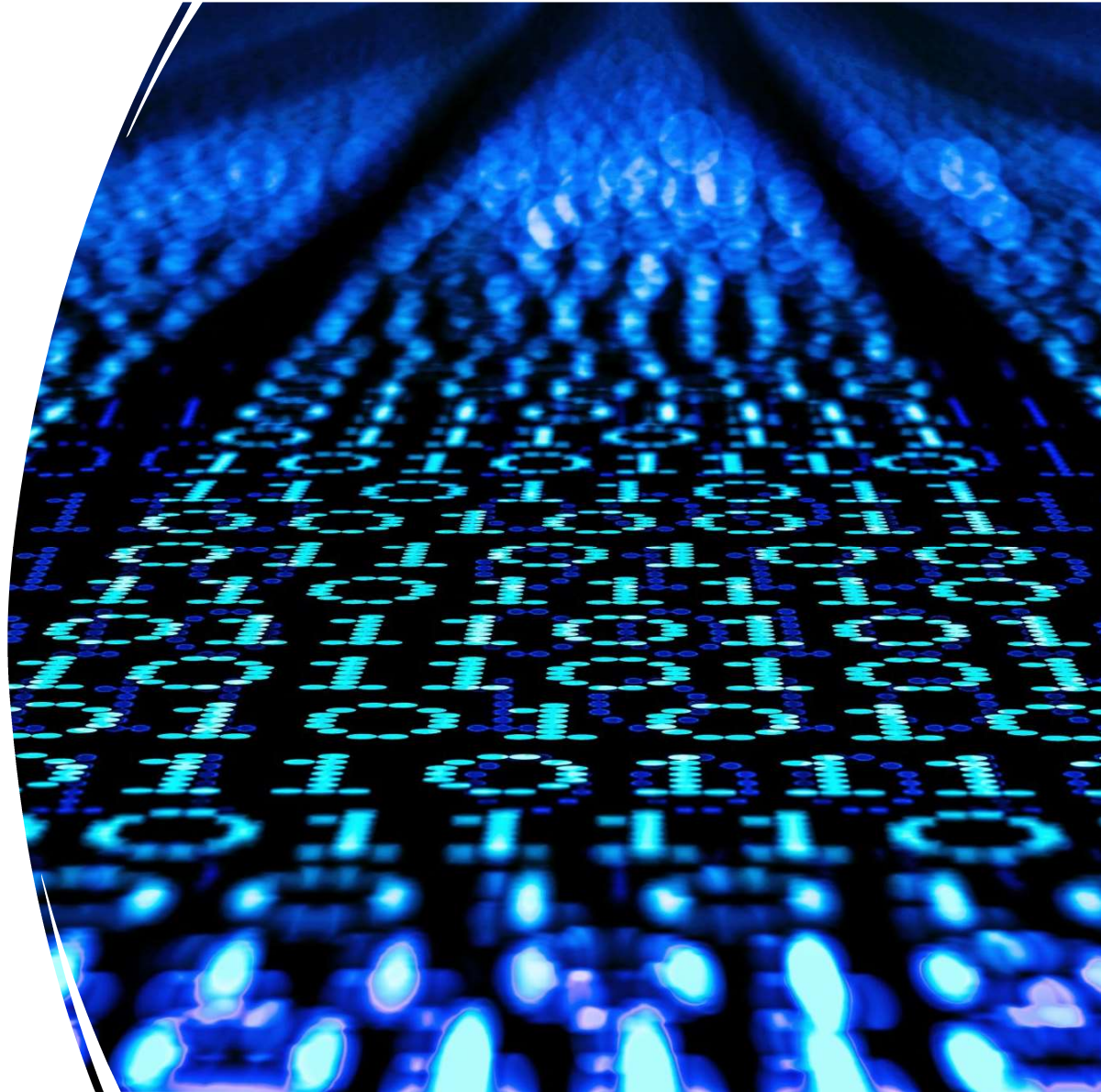
LTERR Network Data Release Policy

- Data and information derived from publicly funded research in the U.S. LTER Network, totally or partially from LTER funds from NSF... [must be].. made available in a community accepted data repository... with as few restrictions as possible, on a nondiscriminatory basis.



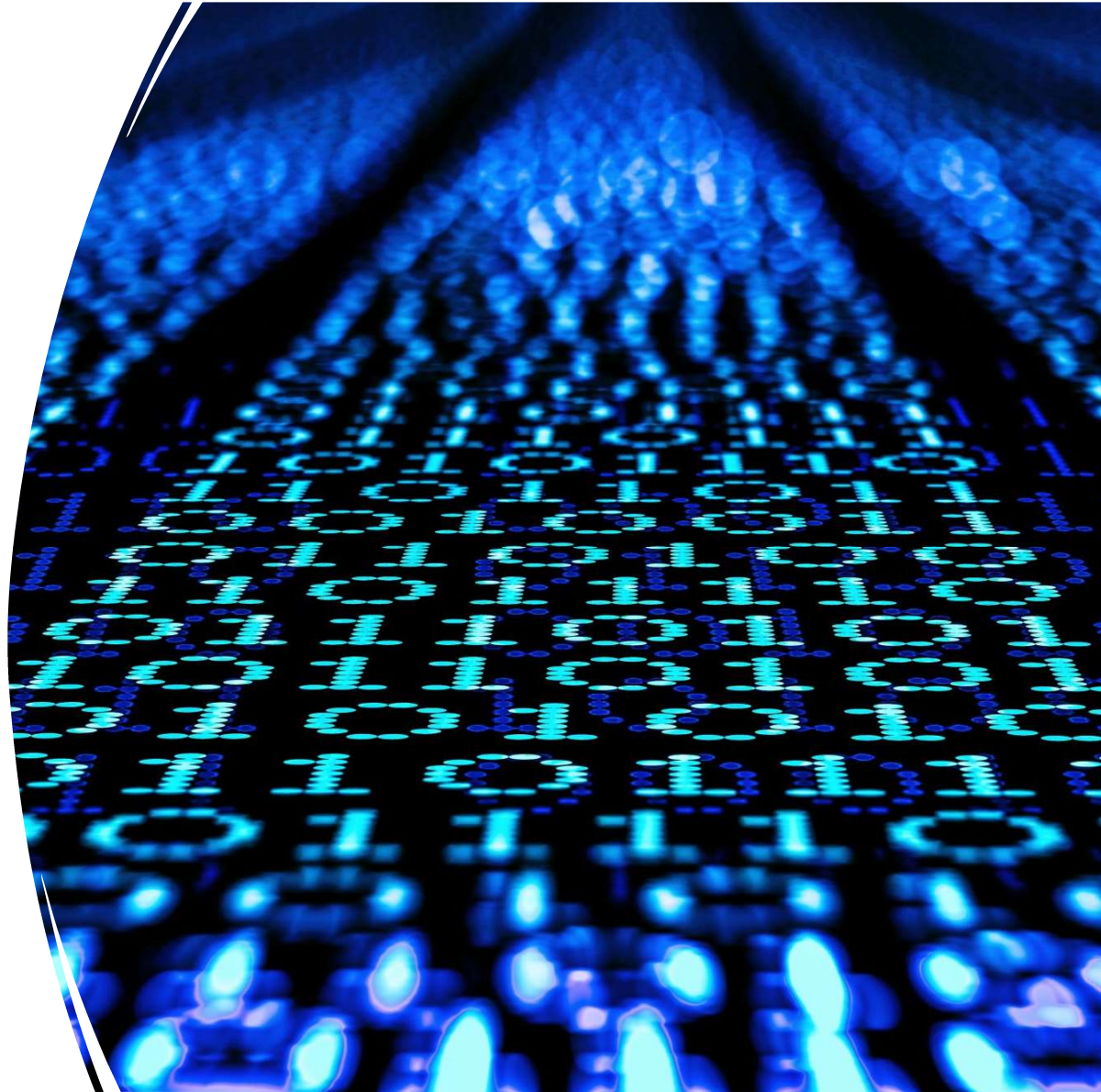
Two Types of LTER Data

- Type I – data are to be released to the general public according to the terms of the general data use agreement **within 2 years from collection** and no later than the publication of the main findings from the dataset
- Type II - data are to be released to restricted audiences according to terms specified by the owners of the data. **Type II data are considered to be exceptional and should be rare in occurrence.**



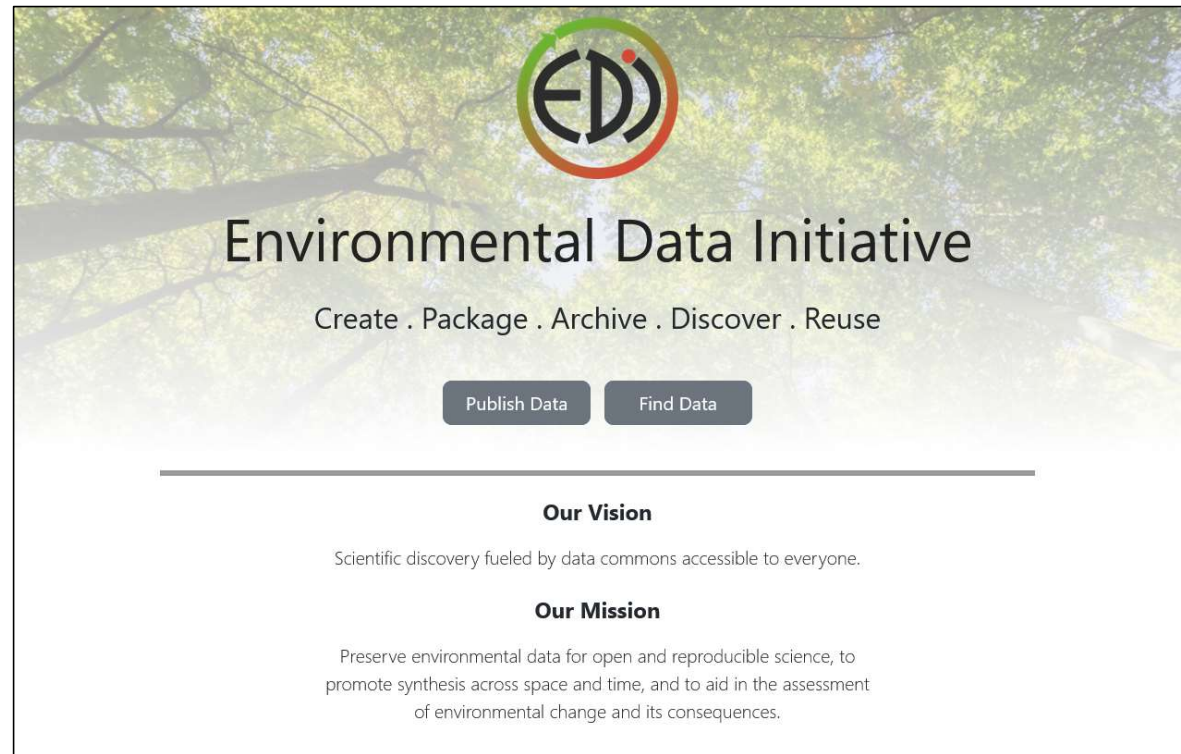
FCE LTER Graduate Student Data

- Submit complete dataset and metadata to FCE Information Manager (IM) before graduation
- Do not wait until the last minute!



The Environmental Data Initiative (EDI) Data Repository

- Funded by NSF
- Provides:
 - Long-term data security
 - Long-term data accessibility
 - Data integrity
 - Data discovery
 - Citable digital object identifiers (DOIs) for datasets



<https://edirepository.org>

EDI Data Portal

- User-friendly interface of EDI repository
- More than 9,900 searchable, unique data packages
- Advanced search functionality

EDI Data Portal

HOME DATA TOOLS HELP LOGIN

enter search terms

▶ ADVANCED SEARCH

Welcome to the EDI Data Portal

How to Submit Data

Data are one of the most valuable products curated by the Environmental Data Initiative (EDI). Data and metadata derived from publicly funded research are made available through this website with as few restrictions as possible, and on a non-discriminatory basis. In return, EDI expects users of data to act ethically by contacting the data provider prior to using it in any published research. In accordance with professional etiquette, data accessed from this website should be cited appropriately when used in a publication. A digital object identifier (DOI) is provided for each dataset to facilitate citation.

The EDI Data Portal contains environmental and ecological data packages contributed by a number of participating organizations. Data providers make every effort to release data in a timely fashion and with attention to accurate, well-designed and well-documented data. To understand data fully, please read the associated metadata and contact data providers if you have any questions. Data may be used in a manner conforming with the license information found in the "Intellectual Rights" section of the data package metadata or defaults to the EDI Data Policy. The Environmental Data Initiative shall not be liable for any damages resulting from misinterpretation or misuse of the data or metadata.

Contributed Data Package Growth

— Unique — All Revisions

Data Packages (Cumulative)

Jan 2013 Apr 2014 Jul 2015 Oct 2016 Jan 2018 Apr 2019 Jul 2020 Oct 2021 Jan 2023 Apr 2024

Contributed Data Packages
Unique: 9946; All Revisions: 36622

Total Data Packages (including
EcoTrends and Landsat)
Unique: 46312; All Revisions: 88391

<https://portal.edirepository.org/nis/home.jsp>

FCE Data Catalog

- Lists FCE datasets published in EDI
- Easily search the catalog from the FCE website!

[Home](#) / [Data](#) / [Core](#)

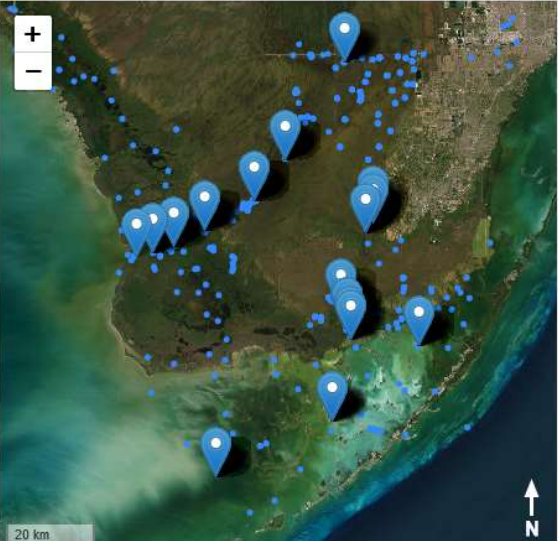
FCE Data Catalog

Search FCE Datasets

Keywords

Originator

ILTER Core Area
All core areas



20 km
10 mi

Leaflet | Powered by Esri | DigitalGlobe, GeoEye, i-cubed, USDA, USGS, AEX, ...

FCE Sites Satellite Sites

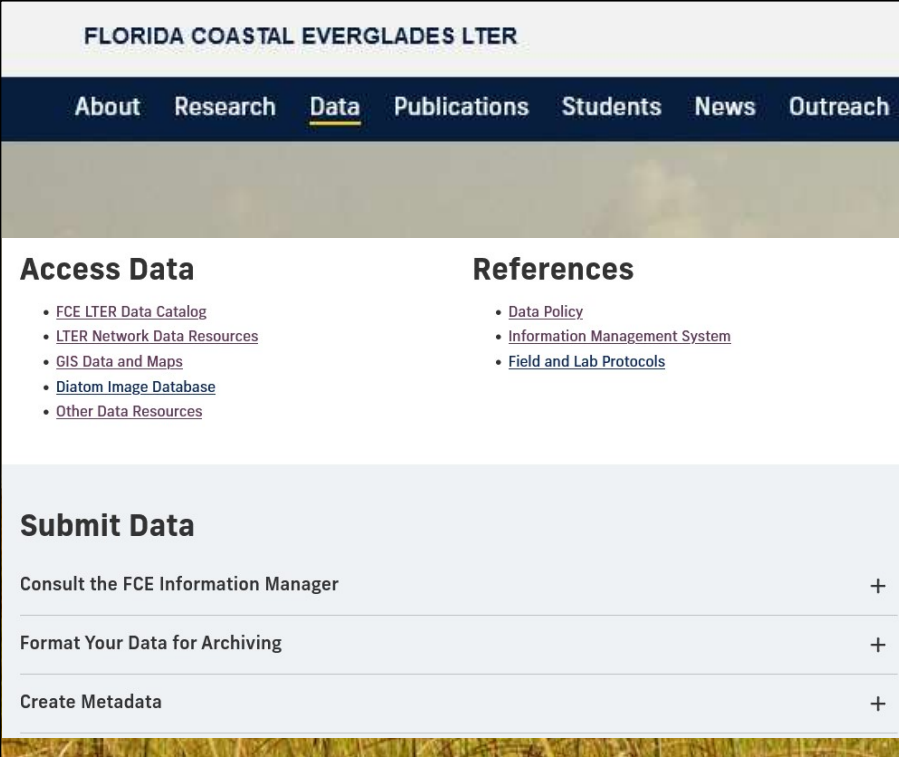
To search for data from specific FCE sites, zoom or pan the to change the area inside a fixed yellow box, which will be included in the search.

Total number of data sets found = 195

<https://fce-iter.fiu.edu/data/core/index.php>

Steps to Publish Data in the EDI Repository

1. Review the FCE Data page!
2. Contact the FCE IM about:
 - Data (tabular, model code, imagery, etc.)
 - Required metadata
 - How to format your dataset
 - Getting an FCE dataset ID
3. Enter data and metadata into ezEML
4. Review package with FCE IM and publish!



FLORIDA COASTAL EVERGLADES LTER

About Research Data Publications Students News Outreach

Access Data

- [FCE LTER Data Catalog](#)
- [LTER Network Data Resources](#)
- [GIS Data and Maps](#)
- [Diatom Image Database](#)
- [Other Data Resources](#)

References

- [Data Policy](#)
- [Information Management System](#)
- [Field and Lab Protocols](#)

Submit Data

Consult the FCE Information Manager +

Format Your Data for Archiving +

Create Metadata +

<https://fcelter.fiu.edu/data>

Best Practices in Data Management

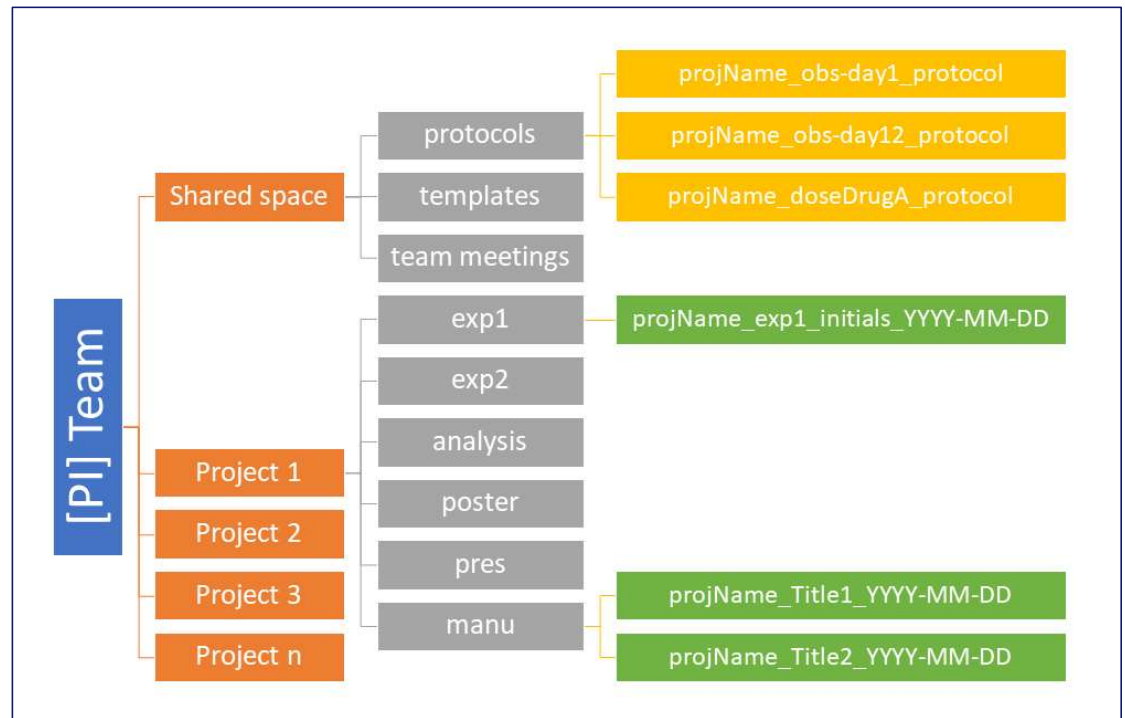
“Well managed data is a benefit to any researcher as it requires less digging to find, less effort to understand, and less processing to prepare for collaboration, reuse, and sharing.” -Briney et al. 2020.



Artwork by @allison_horst (CC BY 4.0)

File Organization

- Use folder structures
- Use consistent file names
- Keep raw data separate from analysis
- Use file versioning



Briney et al. 2020

Backup Your Data

- 3-2-1 rule
 - **Three** copies of the data
 - **Two** geographically separate locations
 - More than **one** type of storage device



Artwork by @allison_horst (CC BY 4.0)

@allison_horst

Write a Living Data Management Plan

- Document important details (e.g. file organization and storage, backup plan, etc.) in one place
- Can be relatively short
- Update as research project evolves



Artwork by @allison_horst (CC BY 4.0)

@allison_horst

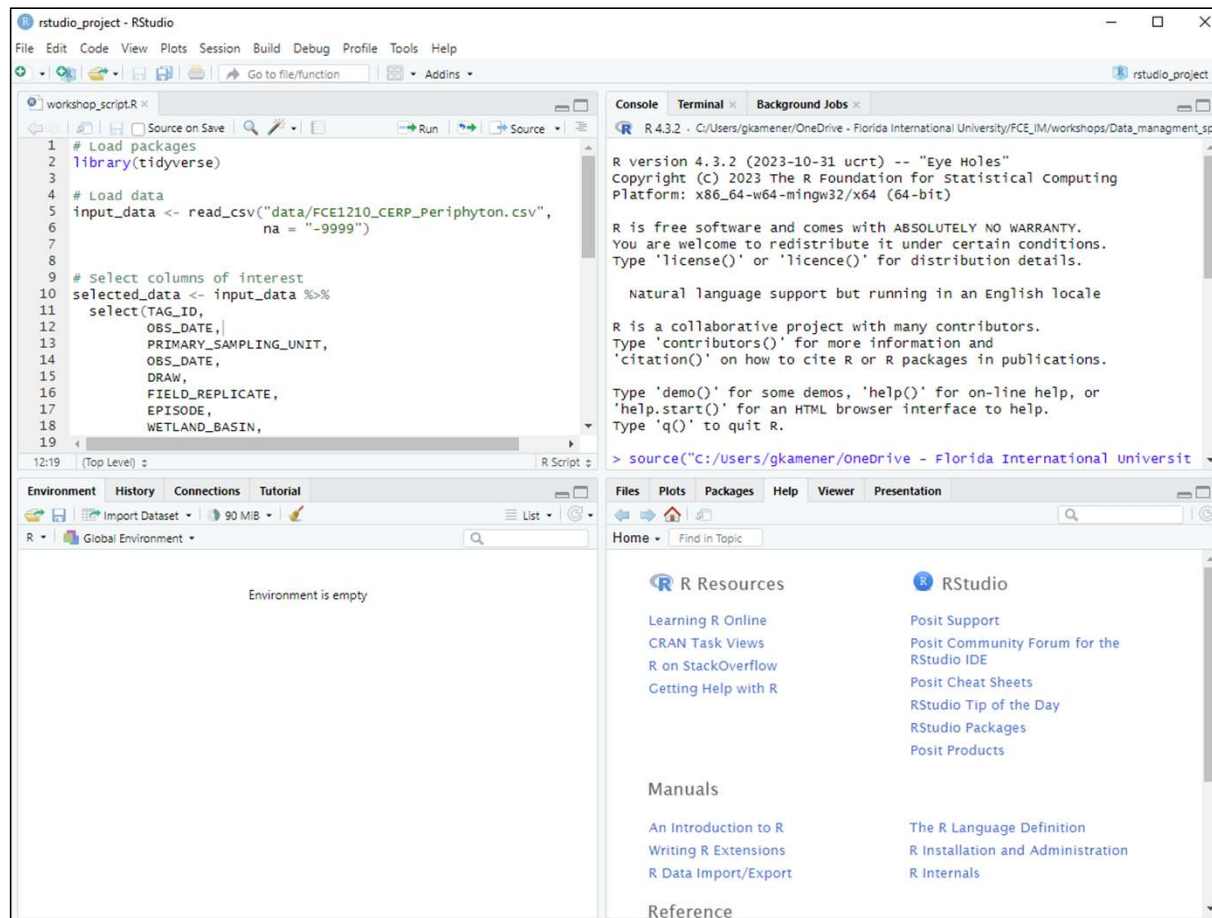
Sharpen Your Data Management Skills

- Briney, K. A., Coates, H. L., & Goben, A. (2020). Foundational practices of research data management. Research Ideas and Outcomes 6: e56508.
<https://doi.org/10.3897/rio.6.e56508>
- Briney, K. (2023). The Research Data Management Workbook. Caltech Library.
<https://doi.org/10.7907/z6czzh-7zx60>



Briney 2023

Project Management in RStudio



RStudio Projects

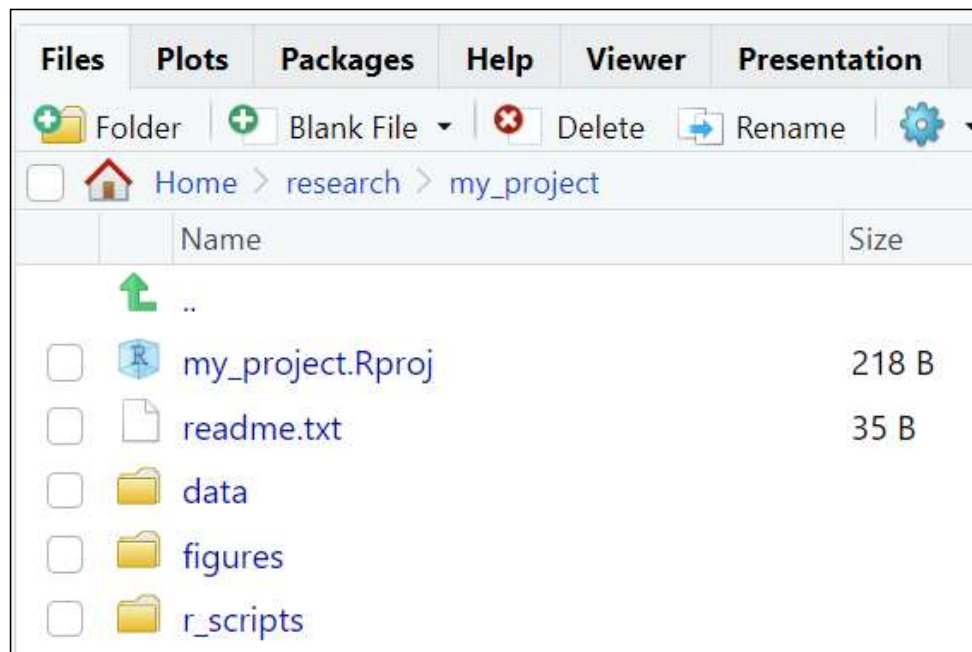
Advantages of RStudio Projects:

- **Automatically set working directory**
- **Portable**
- **Collaborator friendly**
- Supports version control with Git/GitHub
- Can aid reproducibility with renv package

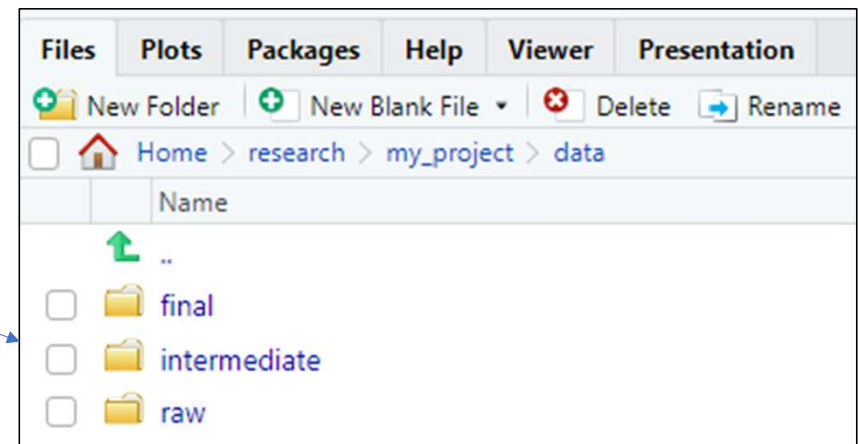
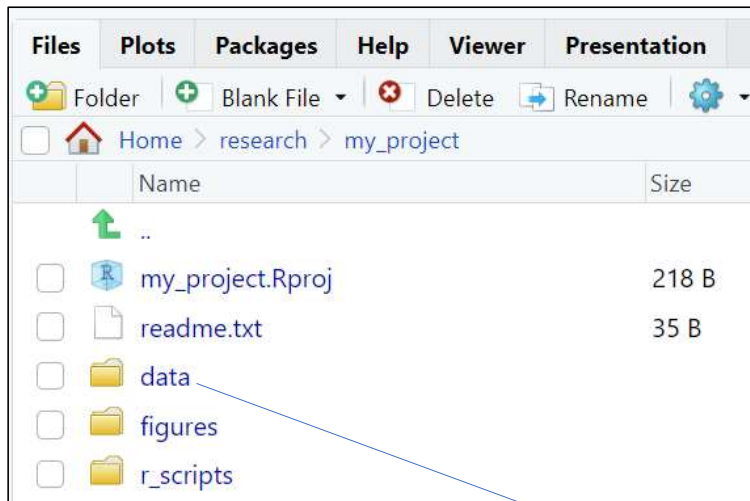
Setting up an RStudio Project

1. Start RStudio.
2. Under the File menu, click on New Project. Choose **New Directory**, then **New Project**.
3. Enter a name for this new folder (or “directory”) and choose a convenient location for it. This will be your working directory.
4. Click on Create Project.
5. Create folders for data (with “raw”, “intermediate”, and “final” subfolders), r scripts, and figures in your working directory.
6. Place raw data files and script files into respective folders.

Organizing Your Working Directory



Organizing Your Working Directory



RStudio Projects: Portability

- R script without RStudio project:

```
df <- read_csv("C:/Users/gkamener/Documents/research/my_project/  
data/raw/my_project_data_GK_2024_11_13.csv")
```

- Often requires specifying working directory or full file path
- Script may break if “my_project” folder is moved or shared

- With RStudio project:

```
df <- read_csv("data/raw/my_project_data_GK_2024_11_13.csv")
```

- Path always relative to “my_project” folder
- Project becomes portable

Formatting Data: Tidy is the Goal

“**TIDY DATA** is a standard way of mapping the meaning of a dataset to its structure.”

—HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

Artwork by @allison_horst (CC BY 4.0)

Formatting Data

- One rectangle of data
 - No blank columns or rows
 - No blank cells
- One data type per column
- One value per cell
- Use simple headers
 - No spaces
 - No special characters



Plot #	Subject ID	Weight1207	Stage1207	Sex
1	1	33.2, 1		m
five	2? Not 100% sure	44.0	2	F
#1	3		three	male



Date	Plot	Subject_ID	Weight	Stage	Sex	Comment
2012-07-01	1	1	33.2	1	m	NA
2012-07-01	5	2	44.0	2	f	Unsure of id
2012-07-01	1	3	-9999.0	3	m	Forgot to measure

Formatting Data

- Format dates as YYYY-MM-DD
- Use consistent categorical variable codes
- Avoid calculations or graphs



Plot #	Subject ID	Weight1207	Stage1207	Sex
1	1	33.2, 1		m
five	2? Not 100% sure	44.0	2	F
#1	3		three	male



Date	Plot	Subject_ID	Weight	Stage	Sex	Comment
2012-07-01	1	1	33.2	1	m	NA
2012-07-01	5	2	44.0	2	f	Unsure of id
2012-07-01	1	3	-9999.0	3	m	Forgot to measure

Formatting Data

- Don't highlight cells or embed comments
- Document any changes you make
- Store metadata in separate sheet or file



Plot #	Subject ID	Weight1207	Stage1207	Sex
1	1	33.2, 1		m
five	2? Not 100% sure	44.0	2	F
#1	3		three	male



Date	Plot	Subject_ID	Weight	Stage	Sex	Comment
2012-07-01	1	1	33.2	1	m	NA
2012-07-01	5	2	44.0	2	f	Unsure of id
2012-07-01	1	3	-9999.0	3	m	Forgot to measure

Metadata

- What are metadata?
 - Metadata = data about data
 - Document **who**, **what**, **why**, **where**, and **when**
- Why use metadata?
 - Keep track of important details about data
 - Required to publish datasets
 - Increase findability and usability of published data

Documenting Metadata

- Dataset metadata should include:
 - Data table information (i.e. data dictionary)
 - Title
 - Abstract
 - Parties responsible for dataset
 - Methods
 - Intellectual rights
 - Keywords
 - Geographic, temporal, and taxonomic coverage
 - Information about non-tabular data (if applicable)
 - Project funding
 - Permits (if applicable)

Documenting Metadata: Data Dictionary

- Describes data table variables (Broman and Woo 2018)
- Create as early as possible!

Column Name	Definition	Variable Type	Units	Precision	Codes	Date Time Format String	Missing Value Code	Missing Value Code Explanation
SITENAME	Name of site	Categorical			SRS2 = Shark River Slough 2			
Date	Date of sample collection	DateTime				YYYY-MM-DD		
Time	Time of sample collection	DateTime				hh:mm	NA	Not recorded
Salinity	Concentration of salinity	Numerical	PSU	0.1			-9999.0	Not recorded
TN	Concentration of total nitrogen	Numerical	micromolePerLiter	0.001			-9999.000	Not recorded
TP	Concentration of total phosphorus	Numerical	micromolePerLiter	0.01			-9999.00	Not recorded
Comment	Field comments	Text					NA	Not recorded

Documenting Metadata

- Title
 - Should be descriptive, including what, where, and when
 - **Do not** use manuscript title
- Abstract
 - Describe who, what, why, where, and when of dataset in more detail
 - **Do not** include results or conclusions from study
- Methods
 - Detail to answer any questions someone might have
 - Include list of cited references (if applicable)
- Intellectual rights
 - FCE uses CC-BY Creative Commons license

Documenting Metadata

- Keywords
 - Source from LTER Controlled Vocabulary* when possible
- Geographic coverage
 - Coordinates in decimal degrees for publishing
- Temporal coverage
 - Documents when data were collected
 - Format as YYYY-MM-DD or YYYY
- Taxonomic coverage
 - Important to check spelling of names

*LTER Controlled Vocabulary link: <https://vocab.lternet.edu/vocab/vocab/index.php>

Documenting Metadata

- Information about non-tabular data entities
 - Includes metadata about model code, geospatial, imagery, or other data entities
- Project funding
 - Information about funding for data collection
- Permits
 - Permits that were required for data collection

Timing and Importance of Quality Metadata

- Important for future you!
- Important for discovery and re-use of published data!
- LTER Network and EDI use Ecological Metadata Language (EML) to document quality metadata

Metadata: Why are they important?

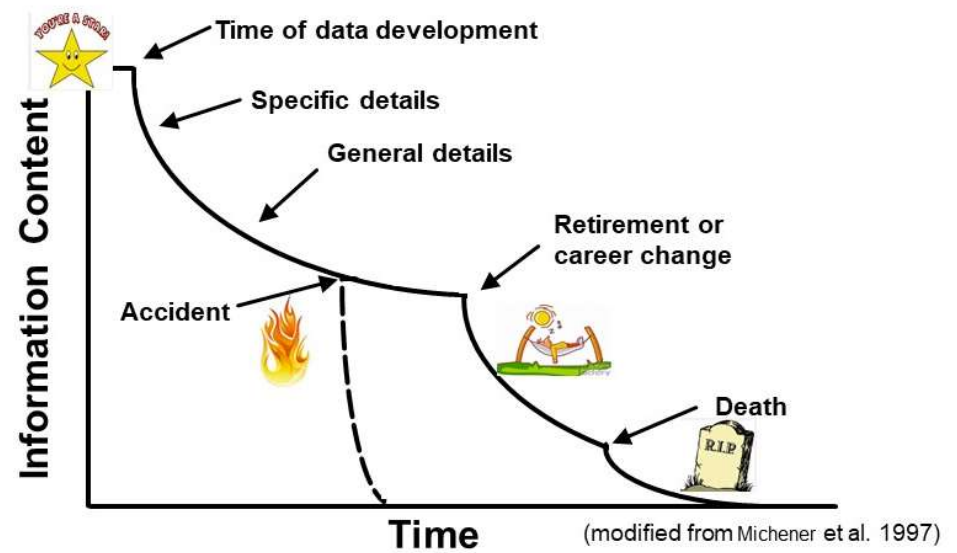


Image courtesy of DataONE

Metadata Creation During Research Life Cycle

- Metadata should be described throughout the research life cycle
 - Describe as much as possible in early stages
 - Regularly add/update throughout life cycle

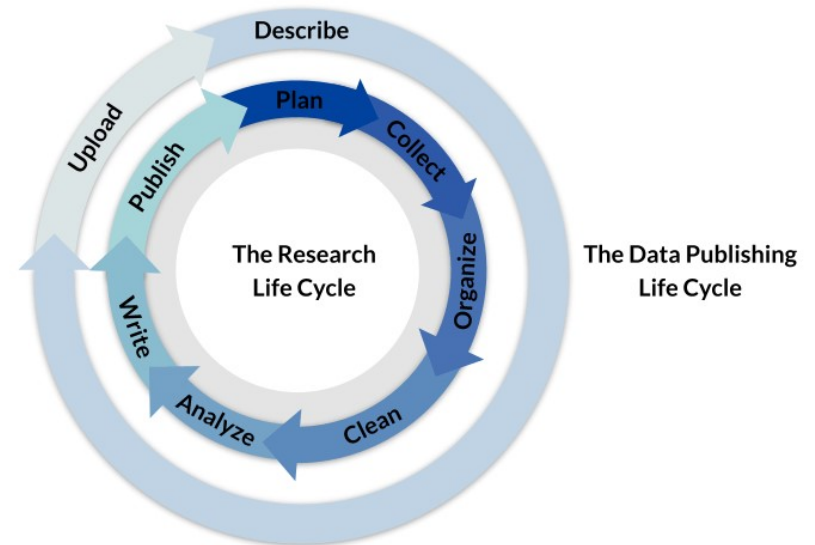
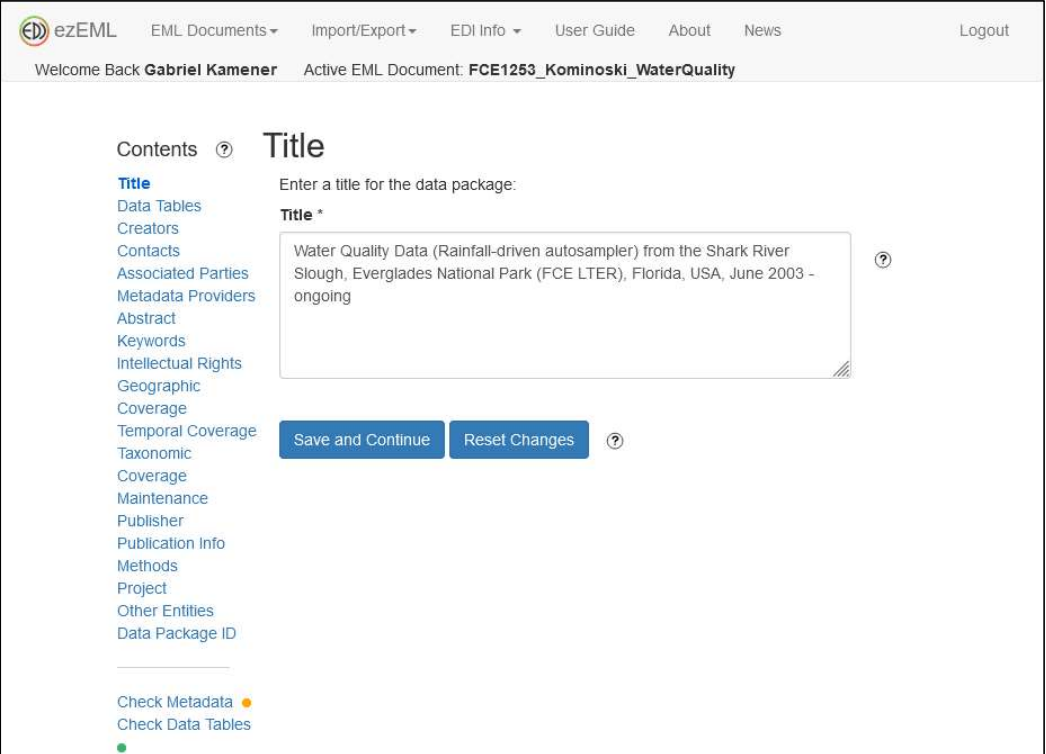


Image courtesy of Environmental data Initiative

ezEML's Role in Data Publishing Life Cycle

- Form-based web interface to create and edit structured metadata in EML
- Can utilize pre-filled FCE templates
- Saves as you go
- Can quality check data and metadata
- Produces data package (data + metadata)

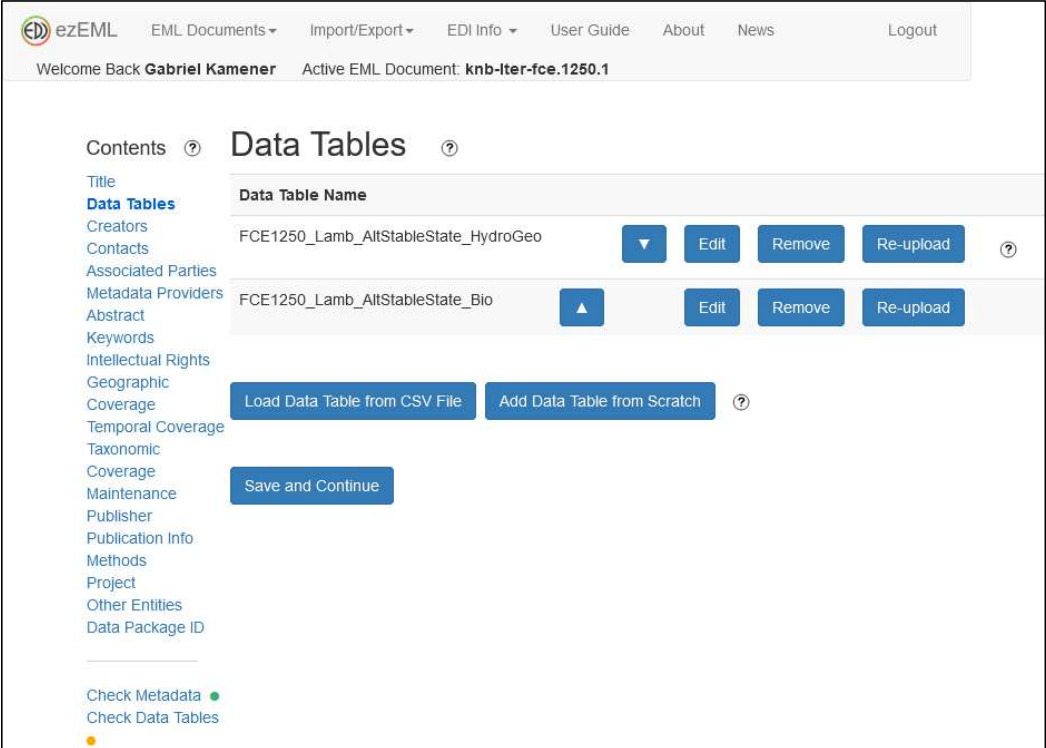


The screenshot displays the ezEML web interface. At the top, there is a navigation bar with the ezEML logo, 'EML Documents', 'Import/Export', 'EDI Info', 'User Guide', 'About', 'News', and 'Logout'. Below this, a welcome message reads 'Welcome Back Gabriel Kamener' and 'Active EML Document: FCE1253_Kominoski_WaterQuality'. The main content area is titled 'Contents' and lists various metadata sections: 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, and Data Package ID. The 'Title' section is currently active, showing a form with the label 'Enter a title for the data package:' and a text input field containing 'Water Quality Data (Rainfall-driven autosampler) from the Shark River Slough, Everglades National Park (FCE LTER), Florida, USA, June 2003 - ongoing'. Below the input field are two buttons: 'Save and Continue' and 'Reset Changes'. At the bottom of the page, there are two status indicators: 'Check Metadata' with a yellow dot and 'Check Data Tables' with a green dot.

<https://ezeml.edirepository.org/eml>

ezEML's Role in Data Publishing Life Cycle

- Can create metadata for tables and other data entities (e.g. model code, imagery, and other nontabular data files)



The screenshot displays the ezEML web interface for an active EML document. The top navigation bar includes the ezEML logo, a user profile for Gabriel Kamener, and various utility links like 'EML Documents', 'Import/Export', 'EDI Info', 'User Guide', 'About', 'News', and 'Logout'. The main content area is titled 'Data Tables' and shows a list of two data tables. The first table is named 'FCE1250_Lamb_AltStableState_HydroGeo' and the second is 'FCE1250_Lamb_AltStableState_Bio'. Each table entry has a dropdown arrow, 'Edit', 'Remove', and 'Re-upload' buttons. Below the list are buttons for 'Load Data Table from CSV File' and 'Add Data Table from Scratch'. A 'Save and Continue' button is also visible. On the left side, a vertical menu lists various metadata categories such as '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', and 'Data Package ID'. At the bottom left, there are status indicators for 'Check Metadata' (green dot) and 'Check Data Tables' (orange dot).

ezEML's Role in Data Publishing Life Cycle

- Quality checks data and metadata
- Provides feedback on potential showstoppers (errors) and things to investigate (warnings)

Contents ⓘ Check Data Table: Results ⓘ

[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](#)

[Check Metadata](#) ●
[Check Data Tables](#) ●

Please note: When data packages are submitted to EDI's data repository, data table error checking is performed there as well. Experienced users of the repository may recognize that the repository's error checking is more permissive than the checking being done here in ezEML. ezEML's error checking is intended to reflect best practices and help data providers minimize the data cleaning burden that will be passed on to consumers of their data.

[Back](#)

Data Table: SRS_Rain_Water_Chemistry

Column: **Date** Type: DATETIME

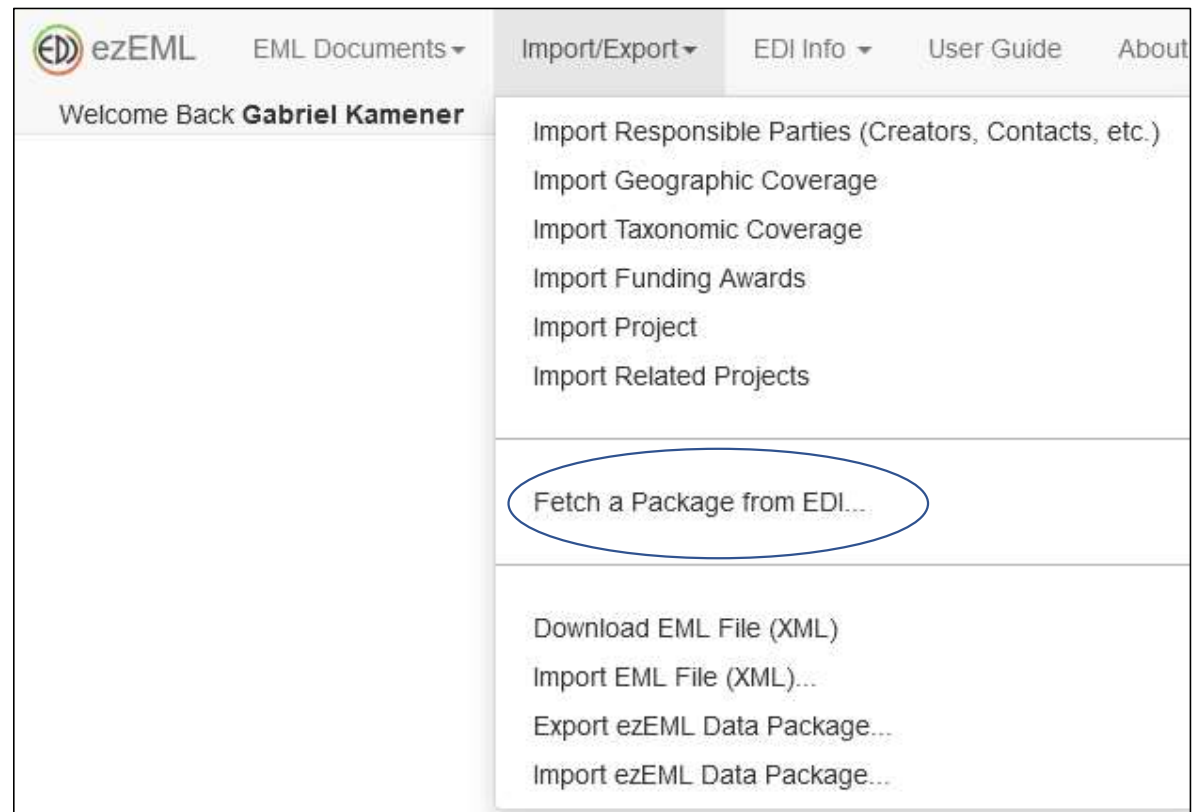
Row	Error	Expected	Found
	The specified DateTime Format String is not supported.	A supported format	mm/dd/yyyy

Column: **Time** Type: DATETIME

Row	Error	Expected	Found
4	DateTime element does not have expected format	hh:mm	9:05
25	DateTime element does not have expected format	hh:mm	6:25
26	DateTime element does not have expected format	hh:mm	6:40
45	DateTime element does not have expected format	hh:mm	9:29

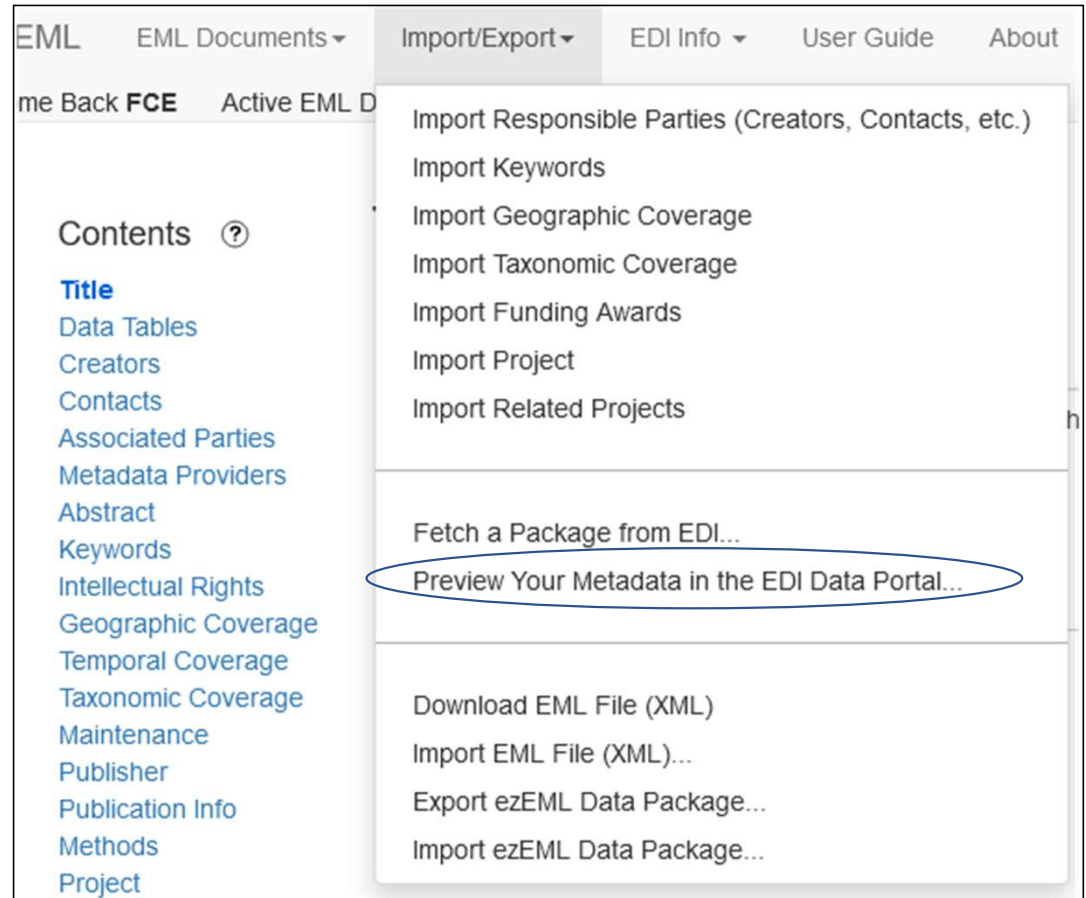
ezEML's Role in Data Publishing Life Cycle

- Can also “fetch” existing packages from EDI to facilitate updates



ezEML's Role in Data Publishing Life Cycle

- Preview how your package will look on the EDI Data Portal



ezEML's Role in Data Publishing Life Cycle

- Preview how your package will look on the EDI Data Portal

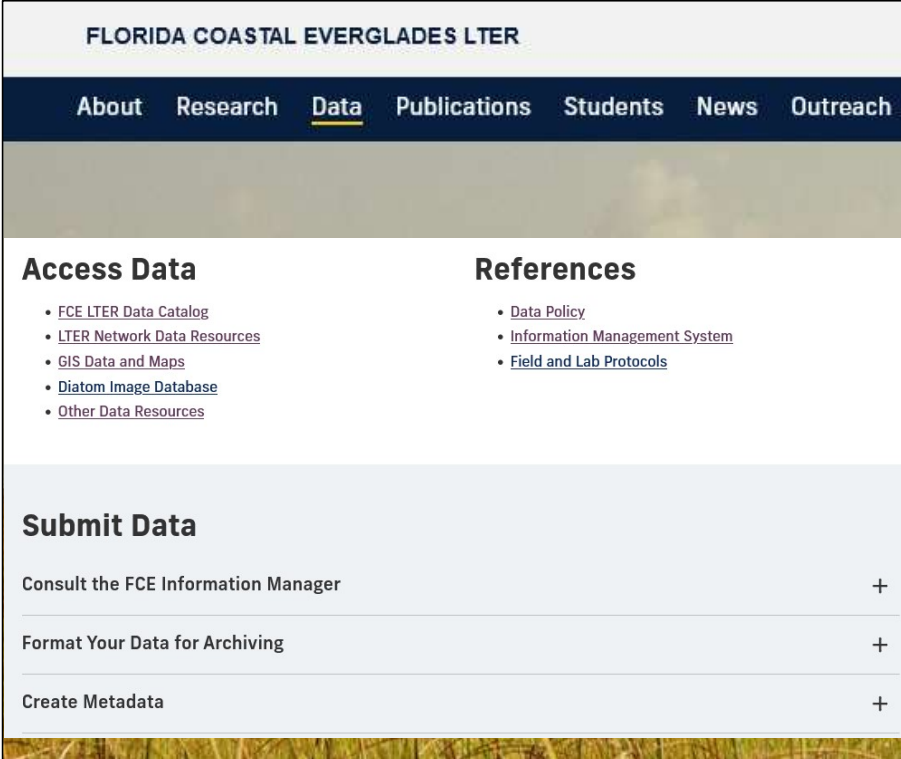


The screenshot displays the EDI Data Portal interface. At the top left is the logo and text "EDI Data Portal". To the right are navigation links: HOME, DATA, TOOLS, HELP, and LOGIN. Below the navigation is a search bar with the placeholder text "enter search terms" and a magnifying glass icon, followed by a link for "ADVANCED SEARCH". The main content area features a blue header for "Data Package Metadata" with a "View Summary" link. The title of the data package is "Water Quality Data (Rainfall-driven autosampler) from the Shark River Slough, Everglades National Park (FCE LTER), Florida, USA, June 2003 - ongoing". Underneath, there is a section for "General Information" which contains a table of metadata.

Data Package:	
Local Identifier:	knb-lter-fce.1253.5
Title:	Water Quality Data (Rainfall-driven autosampler) from the Shark River Slough, Everglades National Park (FCE LTER), Florida, USA, June 2003 - ongoing
Abstract:	Water quality samples are being collected using ISCO autosamplers at all freshwater sites: SRS1a (not active), SRS1c (not active), SRS1d, SRS2, and SRS3. Rain level actuators are used at the sites to trigger water sampling after rain events exceed a given threshold of duration and/or intensity. As currently programmed, when a rain event at a site exceeds the threshold of 2.5 cm per hour, the autosampler at that site collects a 1000mL sample 30 minutes later. The samples are retrieved from the site every 3-4 weeks and analyzed for total phosphorus (TP), total nitrogen (TN), and salinity. Salinity values were not taken consistently from 2000 to mid-2017; those values were replaced by -9999 in the data. See also Shark River Slough precipitation data package (knb-lter-fce.1092) and Shark River Slough extensive water quality data (knb-lter-fce.1072) on the FCE LTER website's data catalog or in the EDI repository.
Publication Date:	2023-09-18

Submitting Your Data to the FCE IM

1. Review the FCE Data page!
2. Contact the FCE IM to review your data and to receive an FCE dataset ID
3. Accept ezEML collaboration invite from FCE IM
4. Enter data and metadata into ezEML
5. Review with FCE IM



The screenshot shows the website for the Florida Coastal Everglades LTER. The header includes the site name and a navigation menu with links for About, Research, Data (underlined), Publications, Students, News, and Outreach. The main content area is divided into two columns: 'Access Data' and 'References'. The 'Access Data' column lists several resources: FCE LTER Data Catalog, LTER Network Data Resources, GIS Data and Maps, Diatom Image Database, and Other Data Resources. The 'References' column lists: Data Policy, Information Management System, and Field and Lab Protocols. Below these columns is a 'Submit Data' section with three expandable items: 'Consult the FCE Information Manager', 'Format Your Data for Archiving', and 'Create Metadata', each with a plus sign to its right.

<https://fcelter.fiu.edu/data>

IM Support at FCE

- Section on FCE Data page!
- Contains IM support information and a growing collection of links to resources on best practices

Access Data <ul style="list-style-type: none">• FCE LTER Data Catalog• LTER Network Data Resources• GIS Data and Maps• Diatom Image Database• Other Data Resources	References <ul style="list-style-type: none">• Data Policy• Information Management System• Field and Lab Protocols
Submit Data	
Consult the FCE Information Manager	+
Format Your Data for Archiving	+
Create Metadata	+
Information Management Support and Best Practices	
Information Management Support	+
Additional Resources	+

<https://fcelter.fiu.edu/data>

IM Support at FCE

- Weekly IM office hours at MMC-CASE 186B, Thursdays 2:00-6:30 pm
- IM support at FCE student Think Tank events
- Available by email or appointment (in-person or Zoom)



Resources

- FCE LTER data page - <https://fcelter.fiu.edu/data>
- [FCE ezEML instructions](#) (PDF file)
- [Tips for submitting FCE data and metadata](#) (PDF file)
- Data Carpentry course episode: [Introduction to R and RStudio](#)
- Briney, K. A., Coates, H. L., & Goben, A. (2020). Foundational practices of research data management. Research Ideas and Outcomes 6: e56508. <https://doi.org/10.3897/rio.6.e56508>
- Briney, K. (2023). The Research Data Management Workbook. Caltech Library. <https://doi.org/10.7907/z6czh-7zx60>
- Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10. <https://doi.org/10.1080/00031305.2017.1375989>
- Google Drive containing this presentation and example metadata documents: <https://tinyurl.com/fce-im-2025>

Questions and Discussion

