

ESIP Biological Data Cluster (BDS) Primer Guide

ESIP Biological Data Cluster

2025-01-15

Table of contents

Preface	4
ESIP BDS Primer Guide Suite of Documents	4
How to contribute	4
Structure	4
1 Introduction	5
2 Provide Context and Understandability to Your Data	7
2.1 Ecological Metadata Language (EML)	7
2.1.1 What Is It?	7
2.1.2 Why?	7
2.1.3 Key Information	7
2.1.4 Top References	8
2.2 ISO 19115	8
2.2.1 What Is It?	8
2.2.2 Why?	8
2.2.3 What?	9
2.2.4 Top References	9
2.3 Minimum Information about any (x) Sequence (MIxS)	9
2.3.1 Who?	9
2.3.2 What is it?	9
2.3.3 Why?	10
2.3.4 Key Information	10
2.3.5 Top References	11
3 INTEGRATE YOUR DATA WITH OTHER DATA	12
3.0.1 Topic: Climate and Forecast	12
3.0.2 Topic: Darwin Core	14
4 Make Your Data Interoperable	17
4.1 Taxonomic Authorities	17
4.1.1 What is it?	17
4.1.2 Why should you use it?	17
4.1.3 Taxonomic Authorities to Know About	17
5 SHARE YOUR DATA ON THE WEB	20

6	Make Your Data Software Ready	27
6.1	Use non-proprietary formats	27
6.1.1	Why?	27
6.1.2	Key Information	27
6.1.3	Top References	27
6.2	Structure tabular data in tidy/long format	28
6.2.1	Why?	28
6.2.2	Key Information	28
6.2.3	Top References	29
6.3	Follow ISO 8601 for dates	29
6.3.1	Why?	29
6.3.2	Key Information	30
6.3.3	Top References	30
6.4	Match scientific names to a taxonomic authority	31
6.4.1	Why?	31
6.4.2	Key Information	31
6.4.3	Top References	31
6.5	Record latitude and longitude in decimal degrees in WGS84	32
6.5.1	Why?	32
6.5.2	Key Information	33
6.5.3	Top References	33
6.6	Use persistent unique identifiers	34
6.6.1	Why?	34
6.6.2	Key Information	34
6.6.3	Top References	35

Preface

ESIP BDS Primer Guide Suite of Documents

The [ESIP Biological Data Standards Cluster](#) formed in 2020 to maximize data relevance and utility for understanding changes in biodiversity over time. To accomplish this the cluster facilitates guidance, best practice documentation, training, and community building for the US biological data community. The first product from this cluster [Biological Data Standards Primer](#), while an easy to digest resource, does not provide the context data managers need to decide which standards to use for the data they are working with. The guides are intended to be a bridge between the full, lengthy standards documentation, and the short primer quick reference. The first document being developed is for the “Make Your Data Software Ready?” section of the primer.

How to contribute

If you would like to suggest changes or additions to the current version of the best practice documents, please use the [GitHub issues](#) to document your request. The current draft can be seen as a rendered webpage [here](#).

Structure

The structure for each section is:

- Value proposition (Why?)
- List / key information (bulleted)
- References list


1 Introduction

This is still a work in progress and only presented here for the purposes of receiving feedback. This message will be removed when the best practices have been officially published.

The diversity of biological data, and (seeming) lack of overarching community standards makes working with biological data challenging. Several standards do exist for biological data, however these different data, metadata, and taxonomic standards are confusing for data managers and data users to navigate. The biological data community in the US could benefit from guidance, best practice documentation, training, and community building. The [ESIP Biological Data Cluster \(BDS\)](#) was formed to tackle these problems.

In short, the will be successful if it:

1. Increase awareness & interest in standards
2. Create unity/shared vision around biological data standards implementation & Provide guidance related to biological standards
3. Provide opportunities for Knowledge sharing & coordination
4. Provide Connectivity across ESIP

We successfully increased awareness and interest in standards through the creation of the primer, officially called, [Biological Observation Data Standardization - A Primer for Data Manager](#). 

Our goals have evolved to work toward creating a sample extension of the cluster's primer in the form of a best practice document(s). The target audience for the best practice document(s) are people who are new to working with biological data standards. The best practice document will provide additional detail on the points listed in the primer, focusing on efficient actions that can be taken to standardize data.

As an example, we brainstormed the “Make your data software ready” section of the primer. Specifically, this includes:

- Use non-proprietary formats
- Structure data in tidy/long format
- Follow ISO 8601 for dates
- Match scientific names to a taxonomic authority
- Record latitude and longitude in decimal degrees in WGS84
- Use globally unique identifiers

The proposed structure for each guide module is:

- Value proposition (Why?)
- List / key information (bulleted)
- References list (max 5)

2 Provide Context and Understandability to Your Data

2.1 Ecological Metadata Language (EML)

2.1.1 What Is It?

EML is a community-developed metadata schema designed for ecological data, which encompasses biological data. EML is normally presented as [Extensible Markup Language \(XML\)](#). An EML instance (XML document) holds metadata to describe one or more data objects. Data tables are the most common, but almost any data object can be accommodated.

2.1.2 Why?

- Provide context to your data and improve reproducibility of the data.
- Can capture linked data relationships within EML (dataset series)
- Standardized representation of information.
- EML was designed for ecological data, which encompasses biological data.
- It's taxonomic fields cover relationships (hierarchies), IDs, and authoritative material

2.1.3 Key Information

- [EML Schema](#)
- Mandatory for LTER, iLTER, OBIS, GBIF, Darwin Core Archive (DwC-A)
- Maintained, and github repo, managed by NCEAS
- Usually, what you would submit to a repository is a “data package” consisting of an EML document and one or more data objects.

2.1.4 Top References

Tools or packages to help write EML:

- For data managers, coders:
 - [EML-R package](#)
 - [Postgresql database with fields compatible with EML](#)
 - [R-code for generating EML from LTER-metabase \(built on EML-R package\)](#)
 - [EMLAssemblyline \(built on EML-R package\)](#)
- For scientists or those not inclined to write scripts
 - [ezEML](#)

2.2 ISO 19115

2.2.1 What Is It?

Content standard for describing geographic data sponsored by the International Standards Organization (ISO). At its most basic, it is written in narrative form with class diagrams. There are many implementations and extensions (e.g., <https://www.dcc.ac.uk/resources/metadata-standards/iso-19115>).

2.2.2 Why?

- Provide context to your data (biological data is inherently ‘geographic’)
- Standardized representation of information
- **Mandated** by some US federal agencies, including NOAA, NASA, and USGS
- Can be used at different granularities, used to describe data packages or collections, as well as at a dataset level (?): content standard vs collection standard?

2.2.3 What?

- Evolved from the need for to to harmonize the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) with other formal and defacto standards that support the documentation of geospatial data and services.
- Many variations including 19115, 19115-1, 19115-2
- From [NCEI](#):
 - ISO 19115 Geographic information – Metadata: The ISO standard for documenting geospatial data.
 - ISO 19115-2 Geographic information – Metadata – Part 2: Extensions for imagery and gridded data: An extension of ISO 19115 used to document information about imagery, gridded data, and remotely sensed data. The root of ISO 19115 metadata records will change from MD_Metadata to MI_Metadata when using ISO 19115-2.
- Usurped FGDC CSDGM - all users encouraged to migrate to ISO.
- Highly flexible for many uses compared FGDC CSDGM, but few required elements leaves room for incomplete metadata

2.2.4 Top References

- [NOAA Workbook for ISO 19115-2](#)
- [How to Convert ISO to EML](#)
- [Work Flow Model](#)
- [mdToolkit](#) - mdEditor is a writer for ISO 19115 metadata which uses mdJSON as an intermediary and mdTranslator allows translation to different metadata formats

2.3 Minimum Information about any (x) Sequence (MIxS)

2.3.1 Who?

This is a standard for molecular data, like DNA and RNA. It is used by molecular biologist and ecologists who generate, manage and archive these type of sequence data.

2.3.2 What is it?

A set of checklists and packages for genomic sequence data.

2.3.3 Why?

- Provide minimal standardized metadata about genetic sequence data
- Agreed upon and published by the Genome Standards Consortium
- Used by the INSDC (DDBJ, EMBL-EBI and NCBI)

2.3.4 Key Information

- MxS (pronounced MIX-ess) is a suite of checklists standards introduced the reporting of a breadth of environment-specific metadata variables to augment the genome-specific checklists.
- Enables mixing and matching of genome checklists and environmental-specific packages.

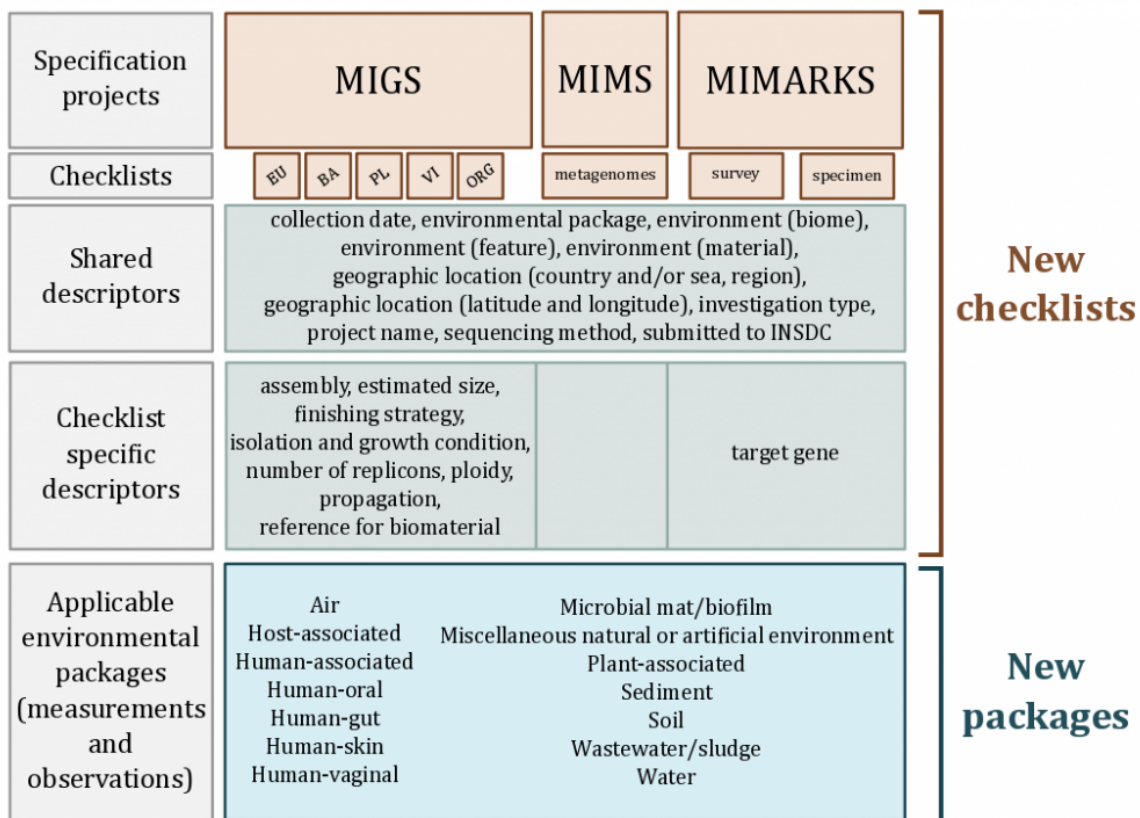


Figure 2.1: MxS Structure

-

2.3.5 Top References

- [MIxS Term Search Tool](#)
- [Genomic Standards Consortium term list](#)
- [Minimum Information about Marker Gene Sequence \(MIMARKS\)](#)
- [MIxS GitHub repo](#)
- [Minimum Information about Sequence Data from the Built Environment \(MIxS-BE\)](#)

3 INTEGRATE YOUR DATA WITH OTHER DATA

- Why?
- Key Information
- Top 5 References

3.0.1 Topic: Climate and Forecast

What is it:

The Climate and Forecast (CF) metadata conventions are designed to promote the processing and sharing of files created with the NetCDF (Network Common Data Form) [API \(Application Programming Interface\)](#). The conventions define metadata that provide a definitive description of what the data in each variable represents, and the spatial and temporal properties of the data. Said more plainly, the conventions explain the extra information (metadata) that clearly describes what each piece of data means and when and where it was collected.

Why?

This enables users of data from different sources to decide which quantities are comparable, and facilitates building applications with powerful extraction, regridding, and display capabilities. The CF convention includes a standard name table, which defines strings that identify physical quantities. CF is well-established, although not perfect for biology. Still, biological standards, and other standards, should consider terms from CF that can be used, before reinventing the wheel.

Key Information (How / What)

- [Version 1.0](#) was released in October 2003, we are now on version 1.12
- CF is a convention built on top of the netCDF standard, and it generalizes and extends the netCDF [COARDS conventions](#).

- Like DwC requires knowledge of the EML standard, CF requires knowledge of others standards. Because CF is a netCDF convention, it assumes the netCDF standard is being followed. And it relies on the UDUNITS system of specifying units (see [Units in CF \(UDUNITS\)](#) below).
- a CF principle is to be self-contained. So for example the CF Standard Names attempt to be as general and well-defined as possible, so the reader does not have to access outside sources to understand the terms.
- To find standard names that describe your data, open up the latest [Standard Name table](#) (as HTML or XML) and search through it for words typically used for your data
- The NERC Vocabulary Server hosts CF, and maintains mappings from CF to other vocabularies. They decompose them into more useful SKOS level semantics. For example: https://vocab.nerc.ac.uk/search_nvs/map/?vocab=P07
 - The NERC Vocabulary server is a platform for searching and browsing vocabularies relating to a variety of domains. NERC hosts a collection of vocabularies, including the Climate and Forecast (CF) vocabulary, and provides an interface for searching for terms across all the hosted vocabularies.

Top 5 References

1. <https://cfconventions.org/>
2. [CF GitHub Discussions](#): announcements, forum for community discussion, questions and answers
3. Current proposals for changing CF (CF GitHub issues): [vocabulary](#) (including standard names), [conventions](#), this [website](#) (including governance)
4. [CF GitHub organisation](#)
5. [CF FAQ](#)
6. [List of software for working with CF](#)
7. [List of Projects and Activities that Use the CF Metadata Conventions](#)
8. [Paper](#) describing the CF data model and reference software
9. Overview of CF basics as a [presentation](#) and [paper](#)
10. <https://www.ogc.org/standard/netcdf/>;
11. <https://gdal.org/en/latest/drivers/vector/netcdf.html>

Repositories that use this data standard

1. NCEI
2. NASA EarthData

3.0.2 Topic: Darwin Core

What is it:

Dublin Core [Link to Dublin Core section in the primer guidelines.] is a set of metadata terms used by libraries to describe physical and digital resources.

Darwin Core (DwC) is a glossary of terms...

The Darwin Core archive is a set of interlinked tables...

Darwin Core is a data standard that offers a stable, straightforward and flexible framework for compiling biodiversity data from varied and variable sources. Darwin Core is an extension of Dublin Core for biodiversity informatics. The standard was originally developed by the Biodiversity Information Standards (TDWG, formerly known as the Taxonomic Database Working Group) community, and it is currently maintained by the Darwin Core Maintenance Interest Group; feedback and participation in development is open to the public.. It includes a glossary of terms intended to facilitate the sharing of information about biological diversity by providing identifiers, labels and definitions, improving data reuse in a variety of contexts. In short, it maps information from multiple sources/institutions in a cohesive way for the broader community. Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information. Taxonomic occurrence data can be standardized to Darwin Core irrespective of the observing method by which the data were collected (e.g., observational data, genomics, imaging, animal tracking). Through the use of common terminology and controlled vocabularies, downstream users can more easily discover, search, evaluate, integrate and compare datasets

- Darwin Core Archive
 - A DwC-Archive includes three things - the data itself, an EML file and a meta.xml file
 - * Some repositories only ask for EML (e.g. EDI)

Why?

Biodiversity data, be it in museum collections, environmental monitoring programs, research programs or civic science projects (e.g. iNaturalist), is collected and managed in many different systems and environments. Additionally, the data is often very heterogeneous within and across these systems, depending on research objectives. Formatting data to the Darwin Core standard can play a fundamental role in facilitating open-access biodiversity data sharing, use and reuse.

Why use DwC in place of OGC, or other standards, like CF? People collect data in all sorts of ways (we list them later in that paragraph), but there are consistent aspects of those: the what, where, and when, and then details of those. It would be great to be able to put all of these disparate collection methods together, and that's what aligning your data allows. The data you collect have details with names you've given them, but there are equivalents with specific names that you can give them to align them.

Key Information (How / What)

In practice, adopting the Darwin Core standard revolves around a standard file format, the Darwin Core Archive (DwC-A). This is the biodiversity informatics data standard that uses the Darwin Core terms to produce a data record for biodiversity data. Essentially, DwC-A is a compressed (.zip) file that contains interconnected text (e.g. csv or tsv) files that enables data publishers to share data using common terminology. These data files are logically arranged in a star-like manner, and typically consist of a core file with one or more extension files, connected through the use of primary and foreign keys.

Aside from text data tables, a DwC-A contains .xml files that facilitate human- and machine-interpretation of the data. A descriptor file (meta.xml) describes the contents of the compressed file, as well as the relationships between the core and any extensions. An eml.xml file describes the datasets contained in the DwC-A.

Advanced Darwin Core

Measurement Or Fact Section? The eMOF can help pull in other vocabularies or standards and cross-link between them.

Top 5 References

- [Darwin Core Quick Reference](#)
- Wiczorek et al., (2012) - Darwin Core: An Evolving Community-Developed Biodiversity Data Standard: <https://doi.org/10.1371/journal.pone.0029715>
- Ecological Metadata Language (*maybe we can point to EML section in Guidelines?)

Repositories that use this data standard

- [Global Biodiversity Information Facility \(GBIF\)](#)
- [Ocean Biodiversity Information System \(OBIS\)](#)
- [The Atlas of Living Australia \(ALA\)](#)

And many more, see this list of [Key projects using Darwin Core](#)

Is there anything missing?

4 Make Your Data Interoperable

4.1 Taxonomic Authorities

4.1.1 What is it?

4.1.2 Why should you use it?

4.1.3 Taxonomic Authorities to Know About

Catalogue of Life (COL)

What is it?

[Catalogue of Life](#) brings together information from taxonomists studying every group of organisms to construct an integrated view of currently accepted species across all taxonomic groups. A list of source datasets can be found [here](#). The primary mission of COL is to deliver a freely accessible list of all species and show which species is referenced by any scientific name, but the tools and services offered by COL also enable taxonomists and other stakeholders to publish and revise species lists for any purpose.

Why should you use it?

Communicate across downstream users which organisms belong to the same group. Taxonomists continue to publish new (and revised) scientific names, which are a fundamental tool to help users to refer to these units of biodiversity, and understand everything that has been learned about its biology, distribution and relevance to mankind. CoL adds persistent identifiers that will enable users to track changes to a scientific name.

How to use it?

Users can browse the [COL Checklist](#), which is updated monthly. COL pulls information from specific data sources, e.g. FishBase (see: <https://www.catalogueoflife.org/data/taxon/49JFH>). CoL also has a Data Pipeline outlining how to best use and manage the taxonomic checklist data held by CoL: <https://www.catalogueoflife.org/about/colpipeline>. COL also has a ChecklistBank API: <https://api.checklistbank.org/>.

ITIS

What is it?

[Integrated Taxonomic Information System](#). Partnership of federal agencies that provides reliable information on taxonomy of plants, animals, fungi and microbes in North America and the world. ITIS has information on over 1.8 million species!

Why should you use it?

ITIS couples each scientific name with a unique taxonomic serial number (TSN) which ensures consistency and accuracy in the naming and classification of species. ITIS includes information on nomenclature, taxonomy, and distribution of species. This is an important tool for identifying and cataloging species and monitoring their populations.

How to use it?

Users can browse on the [ITIS website](#) and through the API.

Paleobiology Database (PBDB)

What is it?

The [Paleobiology Database](#) (PBDB) is an online, expert-curated database that aims to provide taxonomic information for paleobiological taxa of all geological ages. It contains data for almost half a million paleobiological taxa from over 900 different contributors.

Why should you use it?

Checking your paleobiological taxonomic names against the PBDB will ensure the names are up-to-date based on current taxonomic literature. PBDB also provides the taxonomic backbone to the Global Biodiversity Information Facility (GBIF) so aligning your taxonomic names with PBDB will make the process of sharing your data easier.

How to use it?

PBDB can be accessed via their website, a mobile application, and an API. The PBDB website has a [Resources](#) tab where more information about these access points can be found. The same Resources page also includes information on how to contribute taxonomic information to PBDB.

WoRMS:

What is it?

The World Register of Marine Species (WoRMS) is an authoritative and comprehensive list of names of marine organisms. In plain language...

Why should you use it?

One of the reasons WoRMS is highly thought of in the marine community is that the content of WoRMS is curated by taxonomic and thematic experts, not by database managers. Each taxonomic group is represented by an expert who has the authority over the content, and is responsible for controlling the quality of the information. Each of these main taxonomic editors can invite several specialists of smaller groups within their area of responsibility to join them. WoRMS is the taxonomic database used by the Ocean Biodiversity Information System (OBIS), and other important biological initiatives.

How to use it?

WoRMS, and its associated tools, can be explored through your [web browser](#), and through its [API](#) using one of the R packages (e.g. [worrms](#), [taxize](#)) or [python](#) package.

5 SHARE YOUR DATA ON THE WEB

- Why?
- Key Information
- Top 5 References

Topic: Web Services

What is it?

Web services run much of our digital world today. For example, you may be familiar with Amazon Web Services (AWS), which is used for analytics and data service during football games, the olympics, and other sporting events. You can think of a web service as a waiter at a restaurant. You (the user) order food (a request), the waiter (the web service) takes your order to the kitchen (the server or application), and then brings you back your food (the response). This allows different parts of a computer system or different systems altogether to interact without needing to know how each other works internally.

Why should you know about them?

Biological services and platforms like [OBIS](#), GBIF, etc. utilize standard web services to serve data.

These web services are relevant to all sectors collecting and handling biological, biodiversity, and environmental information, including the private, academic research, and government and other operational entities. This is what drives the exchange of scientific information.

Web Services to Know About

Application Program Interfaces (APIs)

Overview: Distributed Model Data Access (<https://www.ncei.noaa.gov/access/distributed-data-access>)

ERDDAP™

What is it?

ERDDAP™ [ur-dap] is a data server that offers users a simple, consistent way to download scientific datasets in common file formats, as well as make graphs and maps. ERDDAP is [used globally](#) to share and integrate disparate data across a range of communities in a standardized way. ERDDAP is often the data service used for oceanographic and atmospheric datasets, but also works great for biological and biodiversity-relevant observations, and for both gridded and tabular data. Data providers can [set up their own ERDDAP server](#) to serve up their data. By using ERDDAP you can incorporate multiple data subsets from different sources into a single workspace. Users can download data from ERDDAP in a multitude of file formats, or as graphs or maps by either using a web page or using the RESTful API in the programming language of their choice.

To facilitate comparisons of data from different datasets, requests and results in ERDDAP use standardized space/time axis, which makes it easier for users to specify data constraints in requests without having to worry about the data format.

Data access: ERDDAP provides a variety of data access methods including via a web browser, OPeNDAP, SOS, WMS, WCS, HTTP, and more.

Data formats: ERDDAP can convert data to various formats such as .csv, .json, .nc, .xls, .mat, .dods, and others (more info [here](#))

Data subsetting: ERDDAP allows users to request a subset of a dataset. It converts the subset to the desired file format available for download.

ERDDAP API: All the information, data and figures made available via ERDDAP are also available via an API. See table dataset API docs [here](#), and for gridded datasets [here](#).

Data search: [Search for data](#) across multiple ERDDAP installations

Additional overall documentation on ERDDAP can be found [here](#). The documentation is in the process of being moved to GitHub [here](#).

Why should you use it?

ERDDAP is free and open source, and makes your data much more accessible. ERDDAP has a [RESTful web service](#) which is designed to be easy for computer programs and scripts to use or interact with.

How to use it?

There are many good tutorials and references on how to use ERDDAP including

- [CoastWatch Training](#) and specifically [ERDDAP basics](#)
- [Awesome ERDDAP](#)

-

Thematic Real-time Environmental Distributed Data Services (**THREDDS**)

What is it?

The Thematic Real-time Environmental Distributed Data Services (THREDDS) server has features and interfaces that makes it easier to explore and use data. Here is a comparison of ERDDAP and THREDDS: <https://jsimkins2.github.io/geog473-673/thredds-and-erddap.html>

[you can't access data by date, you need to know the index number. Less like a database than ERDDAP. Developed a little earlier than ERDDAP. ERDDAP was built to be better than THREDDS from a user perspective. Optimized for data cube, e.g. netCDF, data.

Why should you use it?

How to use it?

Web Map Services

What is it?

A Web Map Service (WMS) is a way to retrieve georegistered map images over the internet to display in applications and web pages. It allows you to view and use maps from different sources that host the maps and data used to create them without needing to download them. The WMS specifications were developed by the [Open Geospatial Consortium \(OGC\)](#) to enable interoperability and use in web browsers, open-source GIS software (ex. [QGIS](#)), and proprietary GIS software (ex. [Esri](#)).

When should you use it?

How to use it?

Provide feedback via github: <https://github.com/ESIPFed/bds-primer-best-practices/issues>

Topic: Web-friendly Standards

What is it?

Web-friendly standards are data standards which facilitate the transfer and handling of data over the Web, its architectures and its services. Data standards that comply with web standards promote online sharing, programmatic discovery, access, and processing of data.

Why should you use them?

Adopting web-friendly standards such as W3C standards, Dublin Core, the DataCite standards, and schema.org helps leverage Web technologies to connect and make discoverable research information across various platforms and disciplines, advancing knowledge.

If data standards aren't web-friendly, data and information will be much harder to "see" via Web services, which are the primary route for global data discovery.

Web-Enabled Standards to Know About

The following standards use Web-friendly and FAIR compliant approaches to promote conformance and interoperability. For example, the use of open and dereferenceable URIs/URLs as persistent identifiers for their properties and types / classes. This means that any web browser or service will be able to act on these standards and link users to the issuing authorities.

[W3C standards](#)

What is it?

"W3C standards define an open web platform for application development. The web has the unprecedented potential to enable developers to build rich interactive experiences, that can be available on any device.

The platform continues to expand, but web users have long ago rallied around HTML as the cornerstone of the web. Many more technologies that W3C and its partners are creating extend the web and give it full strength, including CSS, SVG, WOFF, WebRTC, XML, and a growing variety of APIs."

Why should you use it?

It might not be a question so much of why you should use it as much as that you should become aware that you already use these, and may be able to make your data more FAIR by increasing your awareness of these standards. As the W3C website says, "W3C's proven web standards process is based on fairness, openness, royalty-free, we make the web work, for everyone".

How to use it?

You may use one of the W3C standards listed above to do activities like, rendering web pages, web architecture, and linking data and services. There are a variety of resources available that can be found by searching the internet. We do not currently have a single starting point to recommend here.

Dublin Core

What is it? Dublin Core is a metadata standard of 15 ‘core’ terms originally developed for archives and libraries to describe physical or digital resources. The full set of terms can be found [here](#). Each term is optional and can be used multiple times, as repeated ‘elements.’ All terms are defined as Resource Description Framework (RDF) properties. It has also been formally standardized internationally as ISO 15863.

Darwin Core [[insert link to DwC section of the primer guidelines](#)] is based on Dublin Core and is considered to be an extension for biodiversity information of Dublin Core. For information on further extensions to Darwin Core to capture details of additional information regarding a biological/biodiversity data record, use:

- Extended Measurement Or Facts (eMoF): (<https://rs.obis.org/obis/terms#ExtendedMeasurementOrFact>)

eMoF was developed to be used in combination with the Event Core, but is also compatible with other cores. The eMoF can store measurements or facts related to a biological occurrence, environmental measurements or facts and sampling method attributes. This extension also provides the option to provide identifiers to reference a vocabulary for the measurementType, measurementValue and measurementUnit fields.

- The Humboldt Extension for Ecological Inventories (<https://eco.tdwg.org/>): a standard vocabulary maintained by the [Darwin Core Maintenance Group](#). It is intended to facilitate recording of biodiversity ancillary information in operational settings.

Why should you use it?

How to use it?

Disambiguating the Cores presentation: <https://docs.google.com/presentation/d/1DveHXvY5U5XISl0JocDJ5qUe4PP74dPFSrdryra2m1U/edit#slide=id.p>

DataCite

What is it?

DataCite is an international not-for-profit organization which aims to improve data citation, through helping people use web-enabled standards to connect products and citations. This allows users to i) establish easier access to research data, ii) support data archiving and long-term data preservation, and iii) increase acceptance of research data as legitimate, citable

contributions to a scholarly record, promoting reuse and attribution. DataCite helps mint persistent identifiers, such as digital object identifiers (DOI) to research products, as well as provide recommendations for data citation formats. DOIs are a type of persistent identifier that identify and locate objects in the long-term.

Additional documentation on DataCite can be found at <https://support.datacite.org/docs/datacite-commons>.

Why should you use it?

Assigning a DOI to your research product can enable long-term preservation and accessibility of your research product. Among other things, a main value of DataCite is the connection it creates between users and publishing machinery.

How to use it?

DataCite Consortium members can mint DOIs for their research products, making it part of a larger digital ecosystem and helping connect the research product to researchers through other persistent identifiers e.g., ORCIDs for researchers and ROR for organizations. DataCite has a REST API that enables retrieval, creation, and update of a DataCite DOI metadata record, can be queried: <https://support.datacite.org/reference/introduction>

Schema.org

What is it?

Schema.org is a set of extensible schemas that enables users to embed structured data on their web pages for use and indexation by major search engines. Markup on the webpages or in records helps search engines understand the information presented within and provide richer search results. Schema.org was launched in 2011 by Bing, Google and Yahoo to create and support a common set of schemas for structured data markup on webpages. By using schema.org vocabulary as well as various formats (e.g., JSON-LD) to mark up website content with metadata about itself, making it easier for websites or data records to be searched or indexed.

Schema.org is not a formal standards body, but rather a site where there is documentation on the schemas supported by several major search engines. Schema.org essentially defines a dictionary of terms (types, properties, and enumerated values). So its main hierarchy is formed by a collection of types (or class), each of which has properties that describe the type. Schema.org offers a hierarchically structured set of types which determines which properties can be assigned to a particular type. Essentially this helps search engines look for websites and record and understand the relationships between them.

Documentation on schema.org can be found here: <https://schema.org/docs/documents.html>

Why should you use it?

By implementing schema.org you can make your research more easily and prominently discoverable through major search engines.

How to use it?

You can add schema.org markup to your webpages or records using various online tools, including [Google's Structured Data Markup Helper](#), or by directly adding code to your webpages. You can use different formats to add information to your web content implementing the schema.org vocabulary, such as JSON-LD.

6 Make Your Data Software Ready

6.1 Use non-proprietary formats

6.1.1 Why?

- Allows data to be useful in perpetuity by ensuring data readability and reusability across multiple platforms.
- To align better with the FAIR principles (findability, accessibility, interoperability, reusability)
- Makes data more socially equitable, supporting open science. Proprietary formats can depend on software that require licenses, which not everyone can afford/has access to.

6.1.2 Key Information

- *Non-proprietary formats* are supported by more than one developer and can be accessed with different software systems. For example, comma separated values (CSV) format is becoming an increasingly popular non-proprietary format.
- A *proprietary file format* is a file format of a company, organization, or individual that contains data that is ordered and stored according to a particular encoding-scheme, designed by the company or organization to be secret or with restricted access, such that the decoding and interpretation of this stored data is easily accomplished only with particular software or hardware that the company itself has developed. There may also be costs associated with it and access may be limited. Examples include **Microsoft Excel (xlsx)** and **ESRI shapefiles (shp)**.
- Many applications (e.g. Microsoft Office) allow exporting in multiple formats.

6.1.3 Top References

- Table of commonly used formats for common data types
<https://guides.osu.edu/c.php?g=707751&p=5027409>
- A more detailed table that is specific to US Federal records management
<https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html>

6.2 Structure tabular data in tidy/long format

6.2.1 Why?

This is specifically intended for tabular data

- There is a clear and easy to understand structure that can make your data more machine readable and easier to analyze/visualize
 - Clear structure: one observation per row
 - Data are as atomic as possible (e.g., don't mix types in field)
- In the biological data community, tidy formats are more likely to work with commonly-used software
- Easier to aggregate data across multiple files

6.2.2 Key Information

Example of Wide Format

species	site_01	site_02	site_03
Tilia americana	3	0	3
Pinus strobus	1	1	0

Example of Long Format

species	site	count
Tilia americana	site_01	1
Tilia americana	site_02	3
Tilia americana	site_03	5
Pinus strobus	site_01	0
Pinus strobus	site_02	2
Pinus strobus	site_03	1

- Can be tricky working with multiple column datatypes
- Don't use colors or text formatting in tabular data, and only include column names as metadata. All other notes, definitions, etc. should be in an external metadata file (e.g. data dictionary)

6.2.3 Top References

- Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23.
<https://doi.org/10.18637/jss.v059.i10>
- Data Sharing and Management Snafu in 3 Short Acts (video)
<https://www.youtube.com/watch?v=N2zK3s=Atr-4&t=7s>
- Tips for working with data in BASH
<https://www.datafix.com.au/BASHing/2022-01-12.html>
- Data Organization in Spreadsheets for Ecologists
<https://datacarpentry.org/spreadsheet-ecology-lesson/>
- Cleaning Data and Quality Control
<https://edirepository.org/resources/cleaning-data-and-quality-control#data-table-structure>

6.3 Follow ISO 8601 for dates



Figure 6.1: https://imgs.xkcd.com/comics/iso_8601.png

6.3.1 Why?

- Internationally accepted format used across multiple schemas (e.g. Darwin Core, EML, ISO 19115)
- Removes ambiguity related to timezone, daylight savings time changes, and time of day
- Better software integration of time date/time elements

6.3.2 Key Information

- UTC (AKA Zulu or GMT): Coordinated Universal Time (UTC) is the primary time standard by which the world regulates clocks and time. It is time relative to 0° longitude and is not adjusted for daylight saving time. (from [Wikipedia](#)).
- Conversion to UTC, or between time zones, may depend on daylight savings

Examples: April 3, 2023 standardized to ISO 8601

Description	Written in ISO 8601
Date	2023-04-03
Date and Time with timezone offset	2023-04-03T18:29:38+00:00
Date and Time in UTC	2023-04-03T18:29:38Z
Time Interval in UTC (April 3 - 5, 2023)	2023-04-03T18:29:38Z/2023-04-05T00:29:38Z

Examples: different styles of timezone annotation

Description	Written in ISO 8601
Date	2023-04-03
Date and Time with timezone offset	2023-04-03T18:29:38+00:00
Date and Time in UTC	2023-04-03T18:29:38Z
Time Interval in UTC (April 3 - 5, 2023)	2023-04-03T18:29:38Z/2023-04-05T00:29:38Z

6.3.3 Top References

- ISO 8601 wiki: https://en.wikipedia.org/wiki/ISO_8601
- R package lubridate, OlsonNames()
- Python go-to package, datetime <https://docs.python.org/3/library/datetime.html>
- Article on datetime uncertainty: <https://www.datafix.com.au/BASHing/2020-02-12.html>
- Map of offset from UTC: <https://www.timeanddate.com/time/map/>
- Nice time converter: <https://coastwatch.pfeg.noaa.gov/erddap/convert/time.html>

6.4 Match scientific names to a taxonomic authority

6.4.1 Why?

- To integrate or aggregate datasets, we need a common frame of reference for taxonomic name
- Provides an anchor for the taxonomy as scientific understanding evolves.

6.4.2 Key Information

- Definition: As used here, a taxonomic authority is an online resource that maintains up-to-date species-level classification information and provides persistent identifiers for taxonomic classifications. Example: For the species *Balaenoptera borealis* (Lesson, 1828), the WoRMS taxonomic authority ID link is <https://www.marinespecies.org/aphia.php?p=taxdetails&id=137088> and the LSID is `urn:lsid:marinespecies.org:taxname:137088`.
- Use an existing taxonomic authority (e.g. [World Register of Marine Species](#) , [Integrated Taxonomic Information System](#) , [NCBI taxonomy](#)) and include the authority who manages said information in your metadata
- List of many authorities can be found here: https://resolver.globalnames.org/data_sources
- Make yourself aware of the structure, limits, and history of the authority you are using.
- Adopt standard binomial nomenclature, when possible
- When possible, reference the unique identifier in addition to the nomenclature.
- Always save and document the originally recorded name.
- Put notes about identification uncertainty in a separate column.
- Many authorities have APIs through which you can match names to identifiers.

6.4.3 Top References

- R packages
 - taxize is a taxonomic toolbelt for R. taxize wraps APIs for a large suite of taxonomic databases available on the web
<https://cran.r-project.org/web/packages/taxize/index.html>
 - worrms is an API client for [World Register of Marine Species](#)
http://cran.nexr.com/web/packages/worrms/vignettes/worrms_vignette.html
 - worms: another API client for WoRMS
<https://cran.r-project.org/web/packages/worms/index.html>
 - Ritis: API client for ITIS <<https://cran.r-project.org/web/packages/ritis/>>
- Python packages

– WoRMS API client
<https://github.com/iobis/pyworms>

- Global Names Resolver to compare taxonomic concepts across authorities
<https://resolver.globalnames.org/>
- Article: Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications
<https://doi.org/10.3389/fmars.2021.620702>
- TDWG 2022 Keynote: Richard Pyle, “An Introduction to the Scientific Names of Organisms and the Taxon Concepts they Represent”
<https://www.youtube.com/watch?v=rmTvUUjBxrI>

6.5 Record latitude and longitude in decimal degrees in WGS84

WHAT THE NUMBER OF DIGITS IN YOUR COORDINATES MEANS

LAT/LON PRECISION	MEANING
28°N, 80°W	YOU'RE PROBABLY DOING SOMETHING SPACE-RELATED
28.5°N, 80.6°W	YOU'RE POINTING OUT A SPECIFIC CITY
28.52°N, 80.68°W	YOU'RE POINTING OUT A NEIGHBORHOOD
28.523°N, 80.683°W	YOU'RE POINTING OUT A SPECIFIC SUBURBAN CUL-DE-SAC
28.5234°N, 80.6830°W	YOU'RE POINTING TO A PARTICULAR CORNER OF A HOUSE
28.52345°N, 80.68309°W	YOU'RE POINTING TO A SPECIFIC PERSON IN A ROOM, BUT SINCE YOU DIDN'T INCLUDE DATUM INFORMATION, WE CAN'T TELL WHO
28.523457°N, 80.683094°W	YOU'RE POINTING TO WALDO ON A PAGE
28.52345782°N, 80.683094159°W	"HEY, CHECK OUT THIS SPECIFIC SAND GRAIN!"
28.52345782818284°N, 80.683094159265358°W	EITHER YOU'RE HANDING OUT RAW FLOATING POINT VARIABLES, OR YOU'VE BUILT A DATABASE TO TRACK INDIVIDUAL ATOMS. IN EITHER CASE, PLEASE STOP

Figure 6.2: https://imgs.xkcd.com/comics/coordinate_precision.png

6.5.1 Why?

- Users have to know where you collected this data, which requires a latitude, longitude, reference system and uncertainty.
- Decimal-degrees avoids special symbols (° or ') which is preferable for machine readable formats
- WGS84 is a reference coordinate system that is widely used and incorporated in many GPS units and tools, and recognized as a standard by many government agencies.

6.5.2 Key Information

- If possible, encourage data providers to confirm, and record, the WGS84 datum prior to data collection.
- Understand and report the device/instrument uncertainty associated with your coordinates because it affects the usability of your data.
- Consider including the vertical component (altitude, depth, height off bottom, elevation, etc)
- Generally speaking, **degrees-minutes-seconds** (DMS) can be converted to **decimal-degrees** (DD) by:
 - $DD = d + (min/60) + (sec/3600)$
 - Watch out for mixed formats, like degrees, **decimal-minutes** (DDM).
- Degrees West and South become negative in DD.
 - Values for longitude range from -180 to 180, inclusive.
 - Values for latitude range from -90 to 90, inclusive.

Example Coordinates

Format	Example
Decimal Degrees (DD)	30.50833333
Degrees Minutes Seconds (DMS)	30° 15' 10 N
Degrees Decimal Minutes (DM or DDM)	30° 15.1667 N

6.5.3 Top References

- Existing R/python/ESRI packages/functions
 - R - measurements <https://cran.r-project.org/web/packages/measurements/measurements.pdf>
 - EML <https://eml.ecoinformatics.org/schema/index.html> (find “bounding Coordinates”)
 - CF <https://cfconventions.org/Data/cf-conventions/cf-conventions-1.10/cf-conventions.html#latitude-coordinate>
- Getting lat/lon to decimal degrees
https://ioos.github.io/bio_mobilization_workshop/03-data-cleaning/index.html#getting-latlon-to-decimal-degrees
- Some background on precision
 - <https://www.trekview.org/blog/2021/reading-decimal-gps-coordinates-like-a-computer/#a-note-on-accuracy>

- <https://gis.stackexchange.com/questions/8650/measuring-accuracy-of-latitude-and-longitude>
- DMS to DD calculator
<https://www.fcc.gov/media/radio/dms-decimal> – The three most commonly used datums are WGS84, NAD83, and NAD27. A more complete list can be found here: [https://wiki.gis.com/wiki/index.php/Datum_\(geodesy\)#List_of_Datums](https://wiki.gis.com/wiki/index.php/Datum_(geodesy)#List_of_Datums)

6.6 Use persistent unique identifiers

6.6.1 Why?

- It can be useful to have unique identifiers to unambiguously identify granules of information, e.g. dataset, collection, database, taxonomic concept, etc. This will allow users to precisely refer to the data and allow your data to remain identifiable when aggregated with other datasets.
- To be able to uniquely identify a record in your data system or across data systems. Useful to create relational databases or merge records.
- Although it increases workload, it safeguards against confusion and inefficiency in the future.

6.6.2 Key Information

- There are good reasons to keep an identifier opaque, i.e. it does not indicate anything about the content of information it points to. However, there are also transparent, or semi-opaque identifiers in use that take advantage of semantics to guide humans as well as machines.
- One way to create a unique identifier is concatenation of sampling event, location, time, enumeration of unique observation or event. (e.g. `Station_95_Date_09JAN1997:14:35:00.000`)
- Some prefer using opaque identifiers. (e.g. `10FC9784-B30F-48ED-8DB5-FF65A2A9934E`)
- If there is an existing persistent unique identifier, it's usually a good idea to use it (i.e. when using a taxonomic authority like WoRMS and applying their LSID).
- It is important to manage any identifiers you create, if they are not managed by an authority (e.g. DOIs).
- Important that it be persistent (consider samples possibly moving between institutions)

Examples of PIDs

Type of PID	Use Case	Example
Digital Object Identifier (DOI)	Actionable persistent link for papers, data, and other digital objects	https://doi.org/10.6084/m9.figshare.16806712.v2
International Geo Sample Number (IGSN)	Persistent identifier for physical samples	http://igsn.org/AU1243
Life Science Identifier (LSID)	Persistent structured method for biologically significant data	urn:lsid:marinespecies.org:taxname:218214
Open Researcher and Contributor ID (ORCID)	Persistent actionable link for individuals	https://orcid.org/0000-0002-4391-107X

6.6.3 Top References

- Software and Packages to generate uuids:
 - R - uuid <https://cran.r-project.org/web/packages/uuid/index.html>
 - python - uuid <https://docs.python.org/3/library/uuid.html>
 - <http://guid.one/>
 - <https://guidgenerator.com/>
- Guidance on how to use GUIDs (Globally Unique Identifiers) to meet specific requirements of the biodiversity information community
<http://bioimages.vanderbilt.edu/pages/guid-applicability-final-2011-01.pdf>
- Use of globally unique identifiers (GUIDs) to link herbarium specimen records to physical specimens
<https://bsapubs.onlinelibrary.wiley.com/doi/full/10.1002/aps3.1027>
- A Beginner's Guide to Persistent Identifiers
http://links.gbif.org/persistent_identifiers_guide_en_v1.pdf