ESIP Biological Data Cluster (BDS) Primer Guide

ESIP Biological Data Cluster

2024-03-13

Table of contents

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 Provide Context and Understandability to Your Data

1.1 Ecological Metadata Language (EML)

1.1.1 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.

1.1.2 Why?

- Provide context to your data and improve reproducability 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

1.1.3 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.

1.1.4 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

1.2 ISO 19115

1.2.1 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).

1.2.2 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?

1.2.3 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

1.2.4 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

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

1.3.1 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.

1.3.2 What is it?

A set of checklists and packages for genomic sequence data.

1.3.3 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)

1.3.4 Key Information

- MIxS (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.

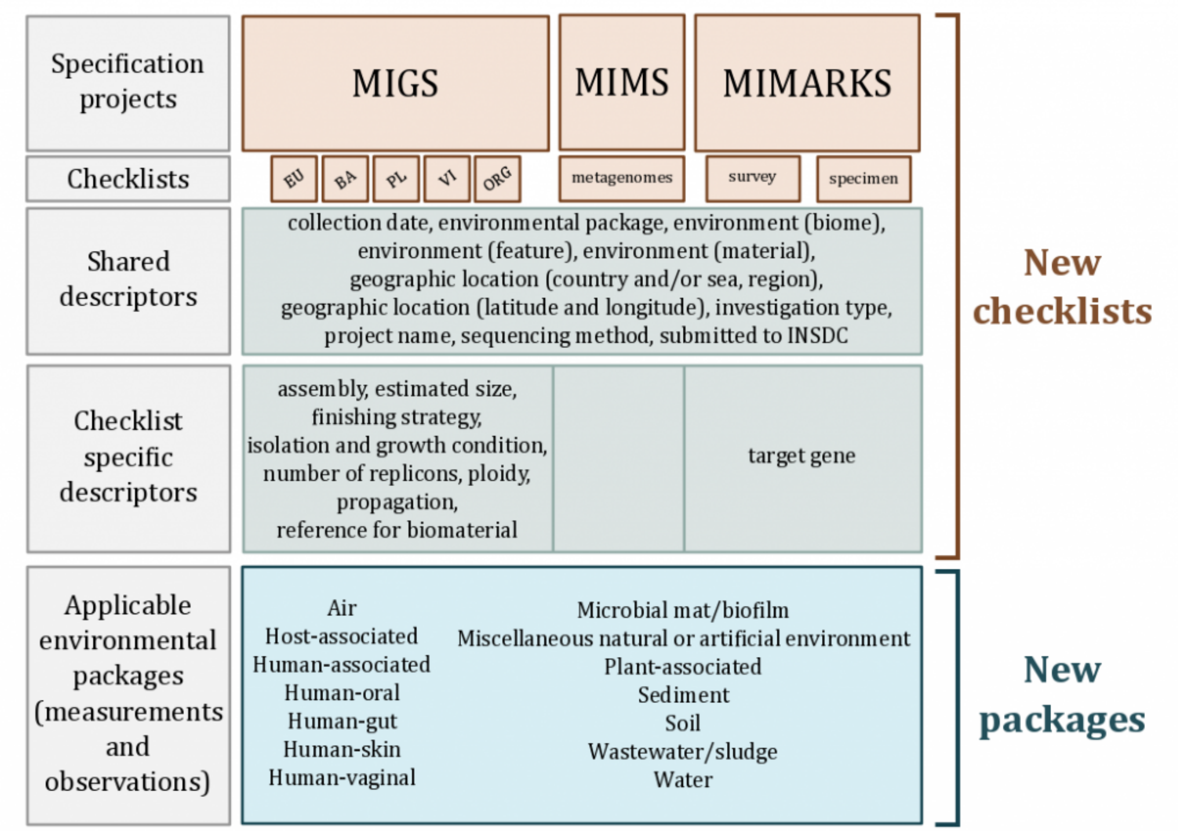

Figure 1.1: MIxS Structure

-

1.3.5 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)

2 Make Your Data Software Ready

2.1 Use non-proprietary formats

2.1.1 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.

2.1.2 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.

2.1.3 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

2.2 Structure tabular data in tidy/long format

2.2.1 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

2.2.2 Key Information

Example of Wide Format

species	site_01	site_02	site_03
Tilia americana	0	4	3
Pinus strobus	3	5	5

Example of Long Format

species	site	count
Tilia americana	site_01	3
Tilia americana	site_02	5
Tilia americana	site_03	4
Pinus strobus	site_01	4
Pinus strobus	site_02	0
Pinus strobus	site_03	2

- 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)

2.2.3 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

2.3 Follow ISO 8601 for dates

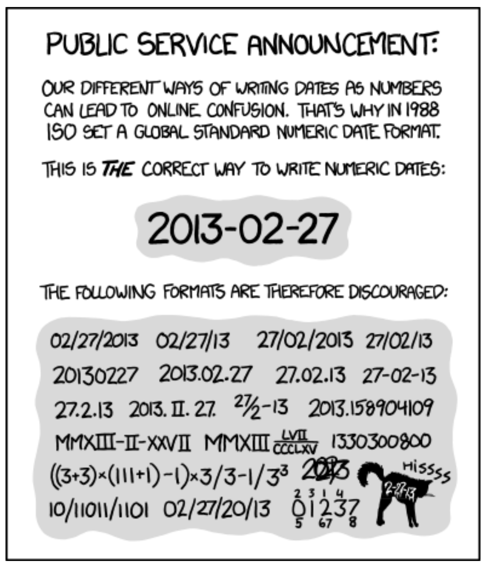

Figure 2.1: https://imgs.xkcd.com/comics/iso_8601.png

2.3.1 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

2.3.2 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

2.3.3 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

2.4 Match scientific names to a taxonomic authority

2.4.1 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.

2.4.2 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.

2.4.3 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

2.5 Record latitude and longitude in decimal degrees in WGS84

Figure 2.2: https://imgs.xkcd.com/comics/coordinate_precision.png

2.5.1 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.

2.5.2 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

2.5.3 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)

2.6 Use persistent unique identifiers

2.6.1 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.

2.6.2 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

2.6.3 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