ESIP Biological Data Standards Cluster Primer: Best Practices

ESIP Biological Data Standards Cluster

2022-12-02

Contents

Section 1

Introduction

This was last updated after the meeting on 2022-12-01, and all material from those brainstorming sessions were incorporated

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 best practice module is:

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

Section 2

PROVIDE CONTEXT AND UNDERSTANDABILITY TO YOUR DATA

- Why?
- Key Information
- Top 5 References

Section 3

INTEGRATE YOUR DATA WITH OTHER DATA

- Why?
- Key Information
- Top 5 References

Section 4

MAKE YOUR DATA INTEROPERABLE

- Why?
- Key Information
- Top 5 References

Section 5

SHARE YOUR DATA ON THE WEB

- Why?
- Key Information
- Top 5 References

Section 6

MAKE YOUR DATA SOFTWARE READY

6.1 Use non-proprietary formats

6.1.1 Why?

- Proprietary data goes against the whole "FAIR" principle.
- Data will be able to be opened by multiple software types
- It could "future proof" data; forward compatibility
- Makes data more socially fair (open data), financial fair
- To align better with the FAIR principles (findability, accessibility, interoperability, reusability)
- Without these, being to use the data as an end user becomes increasingly difficult and will deter people from wanting to use it
- Proprietary software requires licenses, which not everyone can afford/has access to.
- Ensuring data readability across multiple platforms and into the future.
- Contributes to long-term, consistent data collection from multiple data providers

6.1.2 Key Information

- Having open/flat files like csv is useful because they are easily manipulated
- Having your data stored in a relational database can help with ease of conversion to csv/ dwc-a or obis
- How to read said format. Clearly identifying how machines can read format.

6.1.3 Top 5 References

- https://www.go-fair.org/fair-principles/
- https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html
- netCDF spec
- Ascii csv spec (tsv)
- Imagery format specs…

6.1.4 Is anything missing?

- Even though flat files are easy to use, it can be a challenge to know what to place into them. It is also important to have guidance in what to expect as contents in these files (ie unique identifiers, tidy data, basically the other topic).
- What to do when a proprietary format is what your lab uses?!

6.2 Structure data in tidy/long format

6.2.1 Why?

- Easier to analyze.
- more machine readable
- More likely to be usable with other software
- Makes code+analyses more reusable
- Easier to aggregate data across multiple files
- Makes it easier to add a new "MeasurementType" without modifying all past files & software

6.2.2 Key Information

- Example: All scientific names as columns (non-tidy/wide) vs. a single column (tidy/long)
- Darwin Core will require a long format.
- Does tidy/long format (and adding controlled vocab) make the data more machine-readable/actionable?
- Don't want to tell people that they can't use relational databases
- Can be tricky working with multiple column datatypes
- Don't put metadata at the top of your spreadsheet
- Keep it tabular

6.2.3 Top 5 References

- Wickham, H. . (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10

6.2.4 Is anything missing?

- Database schema terminologies

6.3 Follow ISO 8601 for dates

6.3.1 Why?

- Users have to know where you collected this data.
- Consistent across multiple standards - netCDF, DwC, EML
- Well-defined international standard
- Machine readable
- Incorporated into many other standards (e.g. DwC, ISO 19115)
- Clarity of when something is happening across timezones, time changes, etc
- Avoid confusion, internationally understandable, interoperable
- You are able to match information from different data sources using a standardized time stamp.
- Better software integration of time elements (example: time slider on a map)
- Python or R will expect to see time in a particular format.

- Local time can get confusing, must use UTC.

6.3.2 Key Information

- Lat
- Lon
- Vertical component (altitude, depth, height off bottom, etc)
- Table from Wiki showing examples of all possible formats
- Example of Zulu vs local
- Example of date range
- Timezones
- Daylight savings
- Decimal time
- Some datasets (especially using EventCore DwC-A) will be increasingly using date ranges, which seem to break certain packages for

data exploration (YYYY-MM-DDTHH:MM:SS.MMMZ/YYYY-MM-DDTHH:MM:SS.MMMZ)
- As more event-core contributions are made, this might become more of an issue during the manipulation of data; consider installing fix libraries or contacting github communities responsible for issue libraries for fixes in the future.
- Date parts
- Offset from UTC
- Just put your time in this format. 2022-12-01T18:30:35+00:00

6.3.3 Top 5 References

- https://ioos.github.io/bio_mobilization_workshop/03-data-cleaning/index.html#getting-latlon-to-decimal-degrees
- CF documentation - https://cfconventions.org/Data/cf-conventions/cf-conventions-1.10/cf-conventions.html#latitude-coordinate
- EML documentation:
  - High level: https://eml.ecoinformatics.org/schema/index.html (find "bounding Coordinates)
  - spatialRaster:
- ISO 8601 wiki: https://en.wikipedia.org/wiki/ISO_8601
- R package lubridate
- Python go-to package
- https://xkcd.com/1179/
- Article on datteime uncertainty: https://www.datafix.com.au/BASHing/2020-02-12.html

6.3.4 Is anything missing?

- Should datum or projection be prescribed (ie, must it be WGS84?)
- Understanding the metadata schema/data model you're using and how it handles that information.
  - Netcdf: https://cfconventions.org/Data/cf-conventions/cf-conventions-1.10/cf-conventions.html#coordinate-types
  - Iso-19115-2:
  - EML (dataset level, spatialRaster, spatialVector)
- How to convert between deg-min-sec and decimal, e.drounding considerations
- How to convert between projections (if one is prescribed)
- More wide-spread date range support
- More guidelines on how to enter date-ranges
- Can we represent datetime uncertainty?

- Parsers handle dates differently from each other. So sometimes it matters if one uses 2022-12-01 vs 20221201 vs 2022_12_01? (sometimes it depends on the - parser being used to interpret the date)
- What can we do to support data validation tools so data are checked as they are submitted?
- Some spreadsheets don't export dates as ISO

6.4 Match scientific names to a taxonomic authority

6.4.1 Why?

- So researchers are referring to standard names so they can better compare results across studies.
- To more easily handle taxonomic name changes as scientific understanding evolves.
- Efficiently manage global datasets and better aggregation of species information across databases.
- To integrate or aggregate datasets, we have to have a common frame of reference for biological information

6.4.2 Key Information

- Adopting standard binomial nomenclature
- Use an existing taxonomic "backbone" (example: WoRMS (https://www.marinespecies.org/), https://www.itis.gov/)
- Match incoming taxa with the standard taxonomic referencing system.
- Always save and document the original name.
- Identify specific taxonomic authority used for each dataset/record
- Include with the data what "authority" you used for taxonomic

6.4.3 Top 5 References

- OBIS
- https://cran.r-project.org/web/packages/taxize/index.html
- For microbes, either GTDB or NCBI taxonomic naming
- For "plants, animals, fungi, and microbes" Integrated Taxonomic Information System (ITIS)

6.4.4 Is anything missing?

- Unrealistic to mandate single taxonomic database across all fields and all metadata; better to mandate inclusion of DB or reference used in each case
- Is it possible to map from one authority to another via machine readable/AI means?
  - NCBI » GTDB works
  - Would be necessary/have to happen to have this be machine readable?
- How do we define a taxonomic authority ?
  - International recognized, used and actively maintained
- How do we help people go to the "correct" authority?
  - ITIS, WORMs, AOS (https://checklist.americanornithology.org)
  - Is there a 'catalogue' or decision tree for taxonomic authorities?

6.5 Record latitude and longitude in decimal degrees in WGS84

6.5.1 Why?

- Decimal-degrees avoids special symbols (° or ') which is preferable for machine readable formats

6.5.2 Key Information

- All values in degrees-minutes should be converted to decimal-degrees
- Formulas for conversion?
- Examples of other standards?

6.5.3 Top 5 References

- Q: What are existing online tools to make this conversion?
- Existing R/python/ESRI packages/functions

6.5.4 Is anything missing?

- We debated whether or not decimal degrees should be emphasized at all. Is it more important to get the data collectors to use a standard, or show data managers how to convert between standards?

6.6 Use globally unique identifiers

6.6.1 Why?

6.6.2 Key Information

6.6.3 Top 5 References