ESIP Biological Data Cluster (BDS) Primer Guide

ESIP Biological Data Cluster

2023-11-13

# Preface

## ESIP BDS Primer Guide Suite of Documents

The [ESIP Biological Data Standards Cluster](https://wiki.esipfed.org/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](https://doi.org/10.6084/m9.figshare.16806712.v2), 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](https://github.com/ESIPFed/bds-primer-best-practices/issues/new/choose) to document your request. The current draft can be seen as a rendered webpage [here](https://esipfed.github.io/bds-primer-best-practices/introduction.html).

### Structure

The structure for each section is:

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

# 1. PROVIDE CONTEXT AND UNDERSTANDABILITY TO YOUR DATA

* Why?
* Key Information
* Top 5 References

# MAKE YOUR DATA SOFTWARE READY

## Use non-proprietary formats

### 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.

### 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*](https://en.wikipedia.org/wiki/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.

### 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\>

## Structure tabular data in tidy/long format

### 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

### Key Information

Example of Wide Format

| **species** | **site\_01** | **site\_02** | **site\_03** |
| --- | --- | --- | --- |
| Tilia americana | 5 | 2 | 2 |
| Pinus strobus | 4 | 0 | 1 |

Example of Long Format

| **species** | **site** | **count** |
| --- | --- | --- |
| Tilia americana | site\_01 | 2 |
| Tilia americana | site\_02 | 4 |
| Tilia americana | site\_03 | 3 |
| 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)

### 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>

## Follow ISO 8601 for dates

|  |
| --- |
| https://imgs.xkcd.com/comics/iso\_8601.png |

### 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

### 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](https://en.wikipedia.org/wiki/ISO_8601)).
* 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 |

### 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>

## Match scientific names to a taxonomic authority

### 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.

### 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](https://www.marinespecies.org) , [Integrated Taxonomic Information System](https://itis.gov/) , [NCBI taxonomy](https://www.ncbi.nlm.nih.gov/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.

### 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://www.marinespecies.org/)  
    <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>

## Record latitude and longitude in decimal degrees in WGS84

|  |
| --- |
| https://imgs.xkcd.com/comics/coordinate\_precision.png |

### 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.

### 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 |

### 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>

## Use persistent unique identifiers

### 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.

### 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 |

### 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>