

Expanding the network | *Bringing marine research stations and library collections into the data-sharing community*

CESP Workshop Resource List

Table of contents

[What is GBIF](#)

[What is OBIS](#)

[Why Standardize?](#)

[What standards should I use?](#)

[Resources on Data Standards](#)

[Common Standards for Biodiversity Data](#)

[Darwin Core](#)

[International Organization for Standardization \(ISO\)](#)

[Ecological Metadata Language \(EML\)](#)

[Dublin Core \(DCMI\)](#)

[Mapping to Data Standards](#)

[What should I map?](#)

[Data Formatting](#)

[Renaming](#)

[Parsing](#)

[Concatenation](#)

[Standardization](#)

[Validation](#)

[Notes and Catch-Alls](#)

[Digitization](#)

[Resources and Workflows](#)

[A Recommended Workflow for Digitization](#)

[Crowd-Sourcing Digitization Tools](#)

[Tools of the Trade](#)

[A Warning About Spreadsheets](#)

[Tools of the Trade - Taxonomic Names](#)

[Taxonomy - World Register Marine Species \(WoRMS\) Taxon Match](#)

[Taxonomy - Global Names Resolver](#)

[Taxonomy - Catalogue of Life Match](#)

[Taxonomy - GBIF Names Parser](#)

[Tools of the Trade - Coordinates and Geography](#)

[Coordinates - Canadensys Coordinate Converter](#)

[Coordinates - InfoXY](#)

[Coordinates - Map Applications](#)

[Geography - GADM](#)

[Georeferencing - Georeferencing Resources](#)

[Tools of the Trade - Dates](#)

[Dates - Canadensys Date Parser](#)

[Tools of the Trade - More Tools](#)

[OpenRefine](#)

[Publishing](#)

[Registration and Endorsement](#)

[Integrated Publishing Toolkit \(IPT\)](#)

[Darwin Core Archives](#)

[Dataset Classes](#)

[Metadata Resources](#)

[Checklist Data \(Taxon Core\)](#)

[Occurrence Data \(Occurrence Core\)](#)

[Sampling-Event Data \(Event Core\)](#)

[Creative Commons Designation](#)

[What else can I publish?](#)

[Metadata](#)

[Workshop Resources](#)

What is GBIF

The Global Biodiversity Information Facility (GBIF) is an intergovernmental network and research infrastructure, which provides open access to data about all types of life on Earth.

<https://vimeo.com/434831655>

What is OBIS

The Ocean Biodiversity Information System (OBIS) is a global network for marine biodiversity, providing open access to biodiversity and biogeographic data on marine life.

<https://www.youtube.com/watch?v=E6NblAC-1uE>

<https://www.youtube.com/watch?v=mmD-EYNOrFA>

Why Standardize?

A standard is simply an agreed way of doing something. Standards may include conventions, restrictions, rules, requirements, norms and specifications. According to the ISO, standards are the “distilled wisdom of the people with expertise in their subject matter and who know the needs of the organizations they represent”.

(<https://www.iso.org/standards.html>, February 2023)

For those organizations who wish to publish their biodiversity-based data to one of the many biodiversity data aggregators, such as GBIF or OBIS, using the accepted data standards for publication will make those data more:

1. Discoverable in data portals
2. Fit for use, less ambiguous and more understandable by data users
3. Interoperable by APIs and data analysis tools
4. Easily integrated into online resources, including data portals
5. In accord with the [FAIR](#) data principles.

What standards should I use?

When publishing biodiversity data to an aggregator there are several standards that the publishing community has agreed to use. [The Darwin Core Standard](#) is the primary

standard used. Parts of [Dublin Core](#) are included in Darwin Core. The biodiversity community also uses several [ISO standards](#) for specific types of data, as well as additional standards, such as the [Ecological Metadata Language](#) for metadata.

Resources on Data Standards

**[Biological Observation Data Standardization - A Premier for Data Managers](#)

[Foundation Standards Darwin Core \(GBIF\)](#) (YouTube video)


[Darwin Core: Standardizing and integrating biological observations \(ESIP\)](#) (YouTube video)

Common Standards for Biodiversity Data

Darwin Core

Darwin Core (DwC; doi.org/10.1371/journal.pone.0029715) is a standard for sharing data about the occurrence of life on earth and its associations with the environment. It provides a formal set of terms and definitions to ensure the utility of data to facilitate research, learning and sharing. All Darwin Core terms can be found in the [Darwin Core Quick Reference Guide](#). OBIS and GBIF follow DwC standards and have different terms required when publishing data.

- [Required terms for OBIS](#)
- Required Terms for GBIF:
 - [Occurrence Datasets](#)
 - [Checklists](#)
 - [Sampling Event Datasets](#)

For a quick reference sheet contrasting the GBIF/OBIS required terms for Occurrence and Event tables see:  [GBIF_OBIS Required Terms.pdf](#) The tables are also presented below.

Occurrence Table		
Term	Status in OBIS	Status in GBIF
occurrenceID	required	required
eventDate	required	required

scientificName	required	required
basisOfRecord	required	required
kingdom	recommended	required
decimalLatitude & decimalLongitude	required	strongly recommended
scientificNameID	required	not required, accepted
occurrenceStatus	required	not required, accepted
taxonRank	strongly recommended	strongly recommended
coordinateUncertaintyInMeters	strongly recommended	strongly recommended
individualCount, organismQuantity & organismQuantityType	strongly recommended	strongly recommended
geodeticDatum	recommended	strongly recommended
eventTime	recommended	not required, accepted
countryCode	not required, accepted	strongly recommended
informationWithheld	not required, accepted	not required, accepted
dataGeneralizations	not required, accepted	not required, accepted
country	not required, accepted	not required, accepted

Event Table		
Term	Status in OBIS	Status in GBIF
eventID	required	required
eventDate	required	required
decimalLatitude & decimalLongitude	required	strongly recommended

samplingProtocol	strongly recommended	required
samplingSizeValue & samplingSizeUnit	strongly recommended	required
countryCode	strongly recommended	strongly recommended
parentEventID	strongly recommended	strongly recommended
samplingEffort	strongly recommended	strongly recommended
locationID	strongly recommended	strongly recommended
coordinateUncertaintyInMeters	strongly recommended	strongly recommended
geodeticDatum	recommended	strongly recommended
footprintWKT	recommended	strongly recommended
occurrenceStatus	required in occurrence extension	strongly recommended

International Organization for Standardization (ISO)

[The International Organization for Standardization](#) is a body focussed on the development and growth of global standards for a wide variety of topics and processes. The biodiversity community, and Darwin Core, utilizes several ISO standards, including:

- ISO 8601, the Standard for dates and times: https://en.wikipedia.org/wiki/ISO_8601
- ISO 3166, the Standard for country codes: https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

NOTE: in addition to the ISO 3166, the [Getty Thesaurus of Geographic Names](#) is used to create a comprehensive set of higher geographic names in the Darwin Core Standard.

Ecological Metadata Language (EML)

[EML](#) is a metadata standard developed specifically for earth, environmental, and ecological sciences that was developed and is maintained by [The Knowledge Network for Biocomplexity](#). Metadata files contain information descriptions about a dataset. EML files are written using the XML format. This standard is applied in various ways, but it is used most commonly in data publishing through the use of the Integrated Publishing Toolkit and in the creation of a Darwin Core Archive (both detailed in Publishing below).

- [GBIF Metadata Profile: How-to Guide](#)

Dublin Core (DCMI)

[The Dublin Core Standard](#) originated in 1995 as a means to describe digital or physical resources. The success of this standard, specifically the Dublin Core Metadata Initiative (DCMI), was an inspiration for Darwin Core. Many of the terms defined by the DCMI have been incorporated into the Darwin Core.

- Dublin Core terms:
<https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

Mapping to Data Standards

When preparing a biodiversity dataset for publication to data portals such as GBIF and OBIS, data publishers must match the verbatim field names in the source data file to the names used in the [Darwin Core Standard](#). This may involve changing existing field names to Darwin Core terms, adding additional columns that use [Darwin Core compliant field names](#) or splitting one field into multiple fields.

It is **strongly recommended** that all work to prepare a dataset for publication be performed using a copy of the original data source file. Always preserve the original source separately. Working with data can be messy. Please don't risk damaging, losing or applying irreparable changes to your original data.

Useful Tools:

The Norwegian Node to GBIF created the [DwC Excel Template Generator](#). This tool will automatically generate four different types of Excel spreadsheets: Occurrence Core, Measurement Or Fact, Metadata, and a README.

NOTE: This tool works best if the desired Darwin Core fields are known in advance, although a default template can be generated.

Another tool from Norway is the Excel to [Darwin Core Standard \(DwC\) Tool](#). This is a macro Excel spreadsheet for setup, data entry, data validation and file export to DwC (Darwin Core Standard) files formatted to be used with the IPT Integrated Publishing Toolkit. Create templates for the Sampling-event and Occurrence Cores, as well as the MeasurementsOrFacts, Extended MeasurementsOrFacts (EMoF), and Simple Multimedia extensions.

What should I map?

Not all columns or fields in a given database or source file need to be published. Publishers can choose which data to make available to data aggregators. For example, some spreadsheets may include columns that contain curatorial notes or the locations of specimens within storage facilities. In general, these types of data are not useful to data users building niche models or distribution maps. Thus, these fields do not need to be mapped to any data standard, nor do they need to be made available during the publication process, even if they are present in the source file.

Data Formatting

The considerations below apply to all data publishing workflows, but are written with the assumption that publishers are working with data files, such as spreadsheets, that have been exported from a database application or that serve as the original source files for specific collections or datasets. Any changes made to databases directly may require programming expertise or additional funding.

Renaming

Many datasets do not use field names or column headers that are recognized terms in Darwin Core or other biodiversity data standards. To publish to data aggregators, such as GBIF and OBIS, only columns using Darwin Core terms as headers will be recognized.

Review the [list of Darwin Core terms](#) to determine which Darwin Core terms best match the dataset's verbatim field names. Then, either change the field names in each column or add new columns that use Darwin Core terms.

NOTE: Do not change the fields names in your local database unless changing those names allows users of the database to work more efficiently and effectively. Similarly, when working with a non-database source, such as a spreadsheet, always create a working copy of your source.

Parsing

Some fields contain data that can, and should, be reflected in multiple fields. There may be many reasons for parsing a single field into multiple fields, but the two most important include meeting data standards and to improve the discoverability of these data. For example,

- The verbatim field "date" contains "4 May, 1987"

This field should be mapped as is to `dwc:verbatimEventDate` (because it is a non-standard date format), but it can also be parsed into three additional fields `dwc:day`, `dwc:month` and `dwc:year` to increase discoverability.

Similarly,

- The verbatim field "spp." contains "Monodon monoceros"

In this case, `field:ssp.` should be mapped to `dwc:scientificName` and parsed into `dwc:genus` and `dwc:specificEpithet`.

See Tools of the Trade for some recommended parsing tools.

Concatenation

At times, data that should be contained within a single Darwin Core field is spread across two or more verbatim fields. In these instances, those data should be concatenated into, or appended onto, a single field. For example,

- Field “collector1” contains “A. Benson”
- Field “collector2” contains “E. Lawrence”

Both fields contain valid collector names and should be concatenated into dwc:recordedBy separated by a space-pipe-space (A. Benson | E. Lawrence).

- Field “Area” contains “Cape Rodney-Okakari Point Marine Reserve”
- Field “Location” contains “2nm E from Te Hāwera-a-Maki / Goat Island”
- Field “Quad” contains “17”

Together these fields describe a specific location and may be concatenated into dwc:locality (Cape Rodney-Okakari Point Marine Reserve, 2nm E from Te Hāwera-a-Maki / Goat Island, Quad 17). field:Location may also be mapped to verbatimLocality (2nm E from Te Hāwera-a-Maki / Goat Island).

Standardization

An important part of data publication is the standardization of data in fields that use a standardized vocabulary. This includes, but is not limited to many geographic, taxonomic and temporal fields. For example,

- verbatimDate:4 May, 1987 is standardized as eventDate:1987-05-04
- Country name:United States of America is standardized as country:United States

Validation

Some data need to be validated to confirm their quality and completeness. Taxonomic and place names should all be checked against known authorities to confirm that they are

current, accepted and spelled correctly. Coordinates and georeferences should be validated for precision and accuracy. Dates and the names of people, such as collectors, should be checked against known resources and institutional knowledge. For example,

- The collector name, J. Williams, should be validated to confirm that it was James Williams and not his younger brother Joseph Williams.
- Minnehaha County should be checked for spelling and to confirm that it is actually in South Dakota.
- Coordinates 36.6210339, -121.9045944, must be checked for accuracy and precision to confirm that it is actually the location of the Loeb Laboratory at the Hopkins Marine Station. Additional fields for geodeticDatum, uncertaintyInMeters, and other georeferencing fields should be confirmed in support of the coordinates.

Notes and Catch-Alls

Many databases and spreadsheets contain specific fields for notes about the items or observations in the collection. It is not uncommon for these fields to become catch-all locations for any and all data that do not fit into the common geographic, taxonomic or temporal-focussed fields. Disparate data describing physical traits, habitat, sampling methods, collector notes and the location or condition of items in the collection may be found in a single field.

It is recommended that the data in these fields be parsed into more specific fields in the local database. Not only will data publication be simplified, it will also help local users of the database find what they need. If parsing and the creation of new fields is not feasible, then the publisher should consider whether or not these data fields should be published to a public data portal in their current format. It may be better to withhold the contents of these fields until a more suitable set of fields can be applied to organize these data for research and education purposes.

Digitization

Resources and Workflows

[Integrated Digitized Biollections \(iDigBio\)](#), the National Resource for Advancing Digitization of Biodiversity Collections (ADBC) funded by the National Science Foundation.

[What is iDigBio?](#)

[Digitization Resources](#) - all iDigBio resources

- [**Digitization Avenue](#) - information on the task clusters that enable efficient and effective digitization
- [Digitizing from Source Materials \(Documents\)](#)
- [Workflow Modules and Task Lists](#) - from the Developing Robust Object to Image to Data (DROID) Project
- Nelson G, Paul D, Riccardi G, Mast A (2012) Five task clusters that enable efficient and effective digitization of biological collections. ZooKeys 209: 19-45.
<https://doi.org/10.3897/zookeys.209.3135>

A Recommended Workflow for Digitization

(from Austin Mast, iDigBio Director of Digitization, Workforce Development, and Citizen Science Domain, workshop slide presentation, 2023-03-01)

1. Identify your organization's mission and/or vision
2. Recognize two or more ways that digitization supports #1
3. Recognize one or more ways that you know that you've been successful with each item from #2
4. Write first draft of data requirements, including scope, informed by #1-3
5. Sketch a first draft of a workflow
6. Identify what you need to understand better to improve #4 and 5. Acquire that understanding. Improve #4 and 5 and draft (or adopt others') protocols
7. Pilot workflow and protocols
8. Evaluate output against data requirements and update workflow and protocols
9. Share your workflow and protocols with the community

Crowd-Sourcing Digitization Tools

[Notes for Nature](#)

[Zooniverse](#)

Tools of the Trade

A Warning About Spreadsheets

Spreadsheet applications, such as MS Excel, Google Sheets and OpenOffice, can be very useful, even essential, when engaged in data quality, mobilization and publication. They can be used to organize, filter, and modify data. Some collections use spreadsheets in lieu of a database application to maintain data. Regardless of how the spreadsheet application is used, there are some considerations of which anyone working with data should be aware.

Some applications will attempt to Interpret your data, especially those fields that contain numeric values, such as coordinates, measurements, and dates. If not set up correctly, the application will change those values to something that it interprets , or “thinks”, the data is intended to represent. This can result in irreparable damage to or loss of data, if not caught immediately.

The risk of data corruption and loss can be mitigated with the following recommendations:

- When exporting data out of a database, or when creating a file from scratch, choose to create a CSV or TXT file, preferably using a UTF-8 encoding, if possible.
- Never double click a data file to open it. Always export the file into Excel or another application using the Import Wizard or similar workflow. Doing so should provide the opportunity to import the data as text and avoid any unanticipated interpretation of the data.
- Make certain that the columns and cells are set to Text so that any inputs will be preserved as they are entered and not interpreted. Even the “General” format can cause transformations in the data. When performing cut/paste operations between two or more data files, be certain that the incoming data/cells/columns do not contain different formatting.

- Saving the spreadsheet document as a CSV or TXT makes the file more interoperable than a proprietary file format. Thus, importing a file into another application, such as publishing software, or an online tool, is much more likely to be successful and preserve the data contained within.
- Always maintain a copy of the source data, just in case something unexpected occurs. It is better to have to spend the time to correct and repeat a process than it is to have lost data.

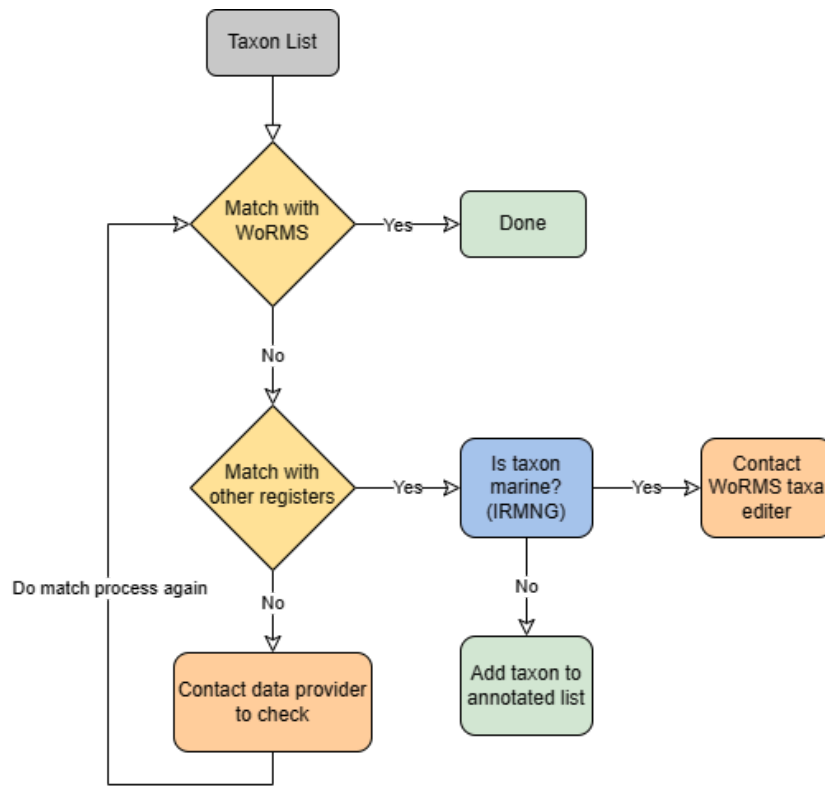
Tools of the Trade - Taxonomic Names

Taxonomy - World Register Marine Species (WoRMS) Taxon Match

<https://www.marinespecies.org/aphia.php?p=match>

WoRMS is an authoritative list of names for marine species, including information on synonyms, as well as providing unique identifiers for species. OBIS uses WoRMS as its taxonomic backbone and the AphiaID from WoRMS is **required** in the scientificNameID field. These unique identifiers can help trace taxonomic changes through time.

If a species cannot be found in WoRMS, we recommend checking if the species is marine using the [Interim Register of Marine and Nonmarine Genera](#). See the flowchart below for how to proceed when matching taxon names.



After obtaining LSIDs from WoRMS, you will need to connect this back to your original dataset. You can do this in R with the [merge](#) function, or in Excel with the [vlookup function](#). You can also complete all taxon matching in R with the [obistools function match_taxa](#). Additional guidance for using the WoRMS taxon match tool is available in documentation from the [Marine Biological Data Mobilization Workshop](#), archived [on Zenodo](#).

Taxonomy - Global Names Resolver

<https://resolver.globalnames.org/>

Resolve lists of scientific names against known sources. This service parses incoming names, executes exact or fuzzy matching as required, and displays a confidence score for each match along with its identifier.

Taxonomy - Catalogue of Life Match

<https://www.catalogueoflife.org/>

[Search Interface](#)

A collaboration bringing together the effort and contributions of taxonomists and informaticians from around the world. COL aims to address the needs of researchers, policy-makers, environmental managers and the wider public for a consistent and up-to-date listing of all the world's known species. COL also supports those who need to manage their own taxonomic information and species lists. COL provides identifiers for taxa, which are similar to, but not the same as those provided by WoRMS.

Taxonomy - GBIF Names Parser

<https://www.gbif.org/tools/name-parser>

This is a simple HTML form to make use of the GBIF name parser. The parser is written in Java and based on regular expressions to dissect name strings into its components. It only keeps name parts required to reconstruct a full three-part name, with an optional subgenus, but ignores additional infraspecific parts such as the subspecies given for varieties. Please see the [API documentation](#) for details.

GBIF Backbone Taxonomy -

<https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c>

Tools of the Trade - Coordinates and Geography

Coordinates - Canadensys Coordinate Converter

<https://data.canadensys.net/tools/coordinates?lang=en>

Use this tool to convert geographic coordinates from DDMSS to decimal degrees.

Coordinates - InfoXY

<http://sblink.cria.org.br/infoxy?criaLANG=en>

A tool to help with the validation of geographic data.

IMPORTANT: Beware the order of the Latitude and Longitude in the U/I.

Coordinates - Map Applications

Google Maps - <https://www.google.com/maps> - click on the map to obtain coordinates from a pop-up at the bottom of the window. NOTE: coordinate precision may be zoom dependent.

OpenStreetMaps - <https://www.openstreetmap.org/#map=4/38.01/-95.84>

Berkeley Mapper - <https://berkeleymapper.berkeley.edu/>

Geography - GADM

<https://gadm.org/>

GADM maps the administrative areas of all countries, at all levels of sub-division. It provides data at high spatial resolutions that includes an extensive set of attributes.

Georeferencing - Georeferencing Resources

Georeferencing Best Practices - <https://docs.gbif.org/georeferencing-best-practices/1.0/en/>

Chapman AD & Wieczorek JR (2020) Georeferencing Best Practices. Copenhagen: GBIF Secretariat. <https://doi.org/10.15468/doc-gg7h-s853>

Georeferencing Quick Reference Guide -

<https://docs.gbif.org/georeferencing-quick-reference-guide/1.0/en/>

Zermoglio PF, Chapman AD, Wieczorek JR, Luna MC & Bloom DA (2020) Georeferencing Quick Reference Guide. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/e09p-h128>

Georeferencing Calculator - <http://georeferencing.org/georefcalculator/gc.html>

Bloom DA, Wieczorek JR & Zermoglio PF (2020) Georeferencing Calculator Manual. Copenhagen: GBIF Secretariat. <https://doi.org/10.35035/gdwq-3v93>

GEOLocate - <https://www.geo-locate.org/>

A platform for georeferencing natural history collections data. Bulk georeferencing is a feature of this tool.

Tools of the Trade - Dates

Dates - Canadensys Date Parser

<https://data.canadensys.net/tools/dates>

A tool to parse dates into their component parts.

Tools of the Trade - More Tools

OpenRefine

<https://openrefine.org/>

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

[User Manual](#)

GBIF Exercise from GBIF's [Biodiversity Data Mobilization Course](#):

- [Instructions](#)
- [Dataset](#)

Publishing

After a dataset has been organized, formatted, and cleaned, it is ready for publication to data portals, such as GBIF and OBIS. The process of publication is more than just sending a CSV or TXT file to a portal; it is the process of making the dataset discoverable and accessible in a standardized format (e.g., Darwin Core) via an access point (e.g., a URL or web address). Some databases and data management systems have a built in means to publish data to URLs, such as Symbiota and the Living Atlases. Most publishers use the Integrated Publishing Toolkit (IPT) to publish to GBIF, OBIS and other data aggregators.

Registration and Endorsement

The first step toward data publication to GBIF and OBIS is to [register the organization or institution as a data publisher](#). Registration begins with the completion of an online form that helps GBIF to learn more about the potential publisher, including its location, collections diversity, provenance of the data shared and available resources for publication. The form also helps GBIF to provide accurate credit and attribution for all data that is shared.

After the form is submitted, the publisher must be endorsed by a GBIF node. The node is usually selected automatically based on the location of the institution. If a node is not available in the country of origin, then the [GBIF Nodes Steering Group](#) will coordinate with the publisher to seek endorsement from an appropriate node. Endorsement ensures that:

1. Published data are relevant to GBIF's scope and objectives
2. Arrangements for data hosting are stable and persistent
3. Data publishing and use are supported by strong national, regional and thematic engagement
4. Data are as open as possible and available for sharing and reuse
5. Data publishers can respond to feedback and improve data quality

Once an endorsement has been made, the institution is welcomed into the global data-sharing community and may publish datasets at will.

Integrated Publishing Toolkit (IPT)

[The IPT](#) is a free, open source software tool used to publish and share biodiversity datasets through the GBIF and OBIS networks. Publishers may choose to [install their own IPT](#) instance or use one of many free IPT installations.

Some IPTs located in North America include:

- [GBIF.us](#) - Any collections from the United States
- [OBIS-USA](#) - Any marine collection from the United States
- [VertNet](#) - Any collection of any taxa from any location on Earth

Other IPTs may be available at private institutions, as well as on any of the [GBIF-hosted IPTs](#). For information about using an existing IPT, please contact [David Bloom](#) (VertNet) or [Abby Benson](#) (OBIS).

[GBIF IPT Users Manual](#)

Darwin Core Archives

The primary purpose of the IPT is to generate Darwin Core Archives. A [Darwin Core Archive \(DwC-A\)](#) is generated for each dataset that is published. The DwC-A is a self-contained set of files that includes a set of text files, in standard comma- or tab-delimited format, including data files, a metadata file (EML.xml) that describes the data files and publishing entity and a descriptor file (meta.xml) that describes how all of the files are organized.

Each DwC-A is published to a unique URL at which data aggregators can access the archive for harvest and inclusion in a data portal.

Dataset Classes

The creation of a Darwin Core Archive begins in the IPT with the selection of one of four data classes or data cores.

Metadata Resources

Institutions can publish datasets describing undigitized resources or other datasets that are not prepared for publication from natural history and other collections. All three other dataset classes include this basic information (metadata), but this 'metadata-only' class offers researchers a valuable tool for discovering and learning about evidence not yet available online. They can also help assess the relative importance and value of undigitized collections and to set priorities for future digitization.

Checklist Data (Taxon Core)

Datasets can provide a catalog or list of named organisms or taxa. Checklists typically categorize information along taxonomic, geographic and thematic lines - often some

combination of the three. These resources may include additional details, such as local species names or specimen citations.

- [Taxon Core data quality requirements](#)
- [DwC-A Checklist Template](#)

Occurrence Data (Occurrence Core)

Datasets published using the Occurrence Core present data that describe the location of individual organisms in time and space. They offer evidence of the occurrence of a species at a particular place on a specified date. Occurrence records may provide only general locality information, such as simply identifying the country, but in many cases more precise locations and geographic coordinates support fine-scale analysis and mapping of species distributions. Occurrence records include descriptions of both physical specimens and human and machine observations.

- [Occurrence Core data quality requirements](#)
- [DwC-A Occurrence Template](#)

Sampling-Event Data (Event Core)

Event Core data often provide greater detail than Taxon or Occurrence data. They offer evidence that a species occurred at a given location and date, but also make it possible to assess community composition for broader taxonomic groups or even the abundance of species at multiple times and places. These sampling-event datasets typically derive from standard protocols for measuring and monitoring biodiversity, such as vegetation transects, bird censuses and freshwater or marine sampling.

The provision of the methods, events and relative abundance of species recorded in a sample make data published with the Event Core the most descriptive of all of the data classes. These datasets improve comparisons with data collected using the same protocols at different times and places.

- [Event Core data quality requirements](#)
- [DwC-A Sampling-Event Template](#)

Creative Commons Designation

An important part of data publishing is the assignment of a [Creative Commons](#) (CC) designation. GBIF requires that all datasets be designated for use with one of three Creative Commons waivers or licenses that state clearly how the published data may be used.

- [CC0](#) - a waiver that states that the data may be used without restrictions
- [CC-BY](#) - a license that states that the data are available for any use if proper attribution is given
- [CC-BY-NC](#) - a license that states that the data may be used for any non-commercial use with proper attribution

These attributions apply only to the data contained within a published dataset. Media related to the occurrence data, such as images, video resources, audio resources and illustrations, are not subject to the CC designation placed on the data. One of the three CC designations may be applied, as well as other CC designations or customized Terms of Use.

To learn more about the application of waivers and licenses, please review [VertNet's Guide to Copyright and Licenses for Dataset Publication](#).

NOTE: In all cases, collections should check with their host institutions and local governments to see if any institutional, cultural, or political restrictions or requirements apply. National laws and regulations will vary. Data publishers are responsible for the adherence to any local or national laws.

What else can I publish?

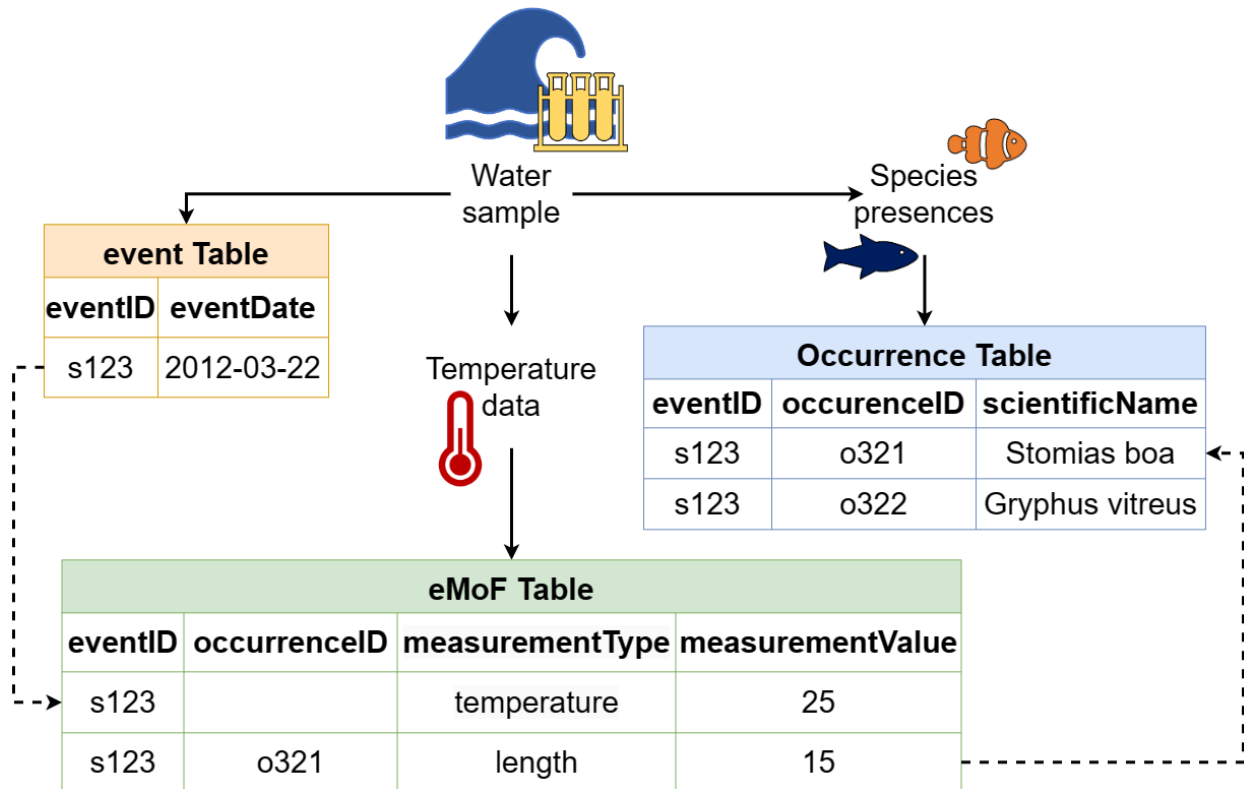
Sometimes, the terms provided by the Simple Darwin Core are not enough to capture the richness or diversity of collections data. For this reason the biodiversity data-sharing community has developed a growing number of extensions to the Darwin Core Standard. These extensions cover a wide range of data types, including media resources, genetic and eDNA data, permitting and sampling data, species distributions, traits and characteristics, and chronometric ages, to name a few.

Many of these extensions can be added to a Darwin Core Archive to extend the information published by the Taxon, Occurrence and Event cores. Even the individual cores themselves

can be published as extensions to other cores. A [list of extensions](#) registered with GBIF is available.

Common Extensions

- [Measurement or Fact](#) - can extend any core and supports the publication of measurements or facts about a specimen or observation. Note that the structure of this table is different from a typical table you may be familiar with. Measurements are “stacked” in rows instead of occurring in separate columns. Measurements are placed in three columns which indicate the type (measurementType, e.g. “length”), the value (measurementValue, e.g. “15”) and the unit (measurementUnit, e.g. “cm”). The eMoF extension below follows the same “long” instead of “wide” structure.
- [Extended Measurement or Facts \(eMOF\)](#) - was developed to be used in combination with the Event Core, but is also compatible with other cores. When used with the Event Core, it allows publishers to create an additional link between the eMOF and the occurrence extension. 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.
- [Audubon Media Description](#) - often referred to as Audubon Core, is a set of vocabularies designed to represent metadata for multimedia resources and collections. These vocabularies represent information that helps to determine whether a particular resource or collection is fit for particular biodiversity science applications before acquiring the media. The vocabularies address the management of the media and collections, descriptions of their content, their taxonomic, geographic, and temporal coverage, and the appropriate ways to retrieve, attribute and reproduce them.
- [DNA Derived Data](#) - captures information related to DNA, including environmental DNA, as an extension to the Occurrence and Event cores. This extension is based on the MlxS extension for Darwin Core (in progress, as of Feb 2023), with additions from [GGBN](#) and [MIQE](#) standards and recommendations. This definition supports the outcomes documented in [Publishing DNA-derived data through biodiversity data platforms](#) (<https://doi.org/10.35035/doc-vf1a-nr22>). This extension is subject to change, and recommended for early adopters who understand that data remapping may be required as things evolve.



Metadata

Documenting the provenance and scope of the dataset is the final part of the preparation of the Darwin Core Archive. When the dataset is published, the metadata will be included as one of the core files (EML.xml) in the archive. Metadata can be entered into the IPT in a variety of ways, but the two most common methods are to:

1. Upload a prepared eml.xml file that contains the appropriate fields and syntax, or,
2. Enter the data manually using the IPT's user interface.

The metadata section of the IPT offers publishers the opportunity to share data in twelve distinct forms

1. Basic Metadata
2. Geographic Coverage
3. Taxonomic Coverage
4. Temporal Coverage
5. Other Keywords

6. Associated Parties
7. Project Data
8. Sampling Methods
9. Citations
10. Collection Data
11. Physical Data
12. Additional Metadata

The GBIF Metadata Template is similar to a manuscript template that makes it easy to author resource metadata. The required fields are all marked clearly. The IPT metadata editor ensures that all mandatory fields have been filled in and that any fields using controlled vocabularies are entered correctly. The IPT also ensures the generated metadata document is valid XML and validates against the GBIF Metadata Profile.

[Metadata Profile How-to Guide](#)

Workshop Resources

Exercise - [Find a Darwin Core Term](#)

Answers - [Find a Darwin Core Term](#)

[Documentation/Cross-reference Template](#)

[Documentation/Cross-reference Example](#)

[Workshop Sample Dataset](#) and [Associated Publication](#)

[Workshop Sample Dataset Cleaned](#)

Exercise - [WoRMS Taxon Match](#) (doc)

Exercise - [WoRMS Taxon Match](#) (Excel)

Answers - [WoRMS Taxon Match](#) (doc)

Answers - [WoRMS Taxon Match](#) (csv)