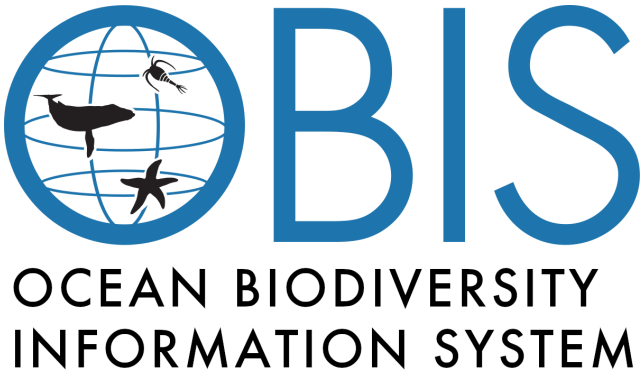
The OBIS manual

21 November, 2023

Contents


OBIS
OCEAN BIODIVERSITY
INFORMATION SYSTEM

Overview

Chapter 1

Introduction

This manual provides an overview on how to contribute data to OBIS and how to acess data from OBIS. It provides guidelines for OBIS nodes and data providers on the OBIS standards and data management best practices to ensure that data published via OBIS are of high quality and follows internationally recognised standards. It also provides guidelines for data users on how to access, process and visualize data from OBIS.

The OBIS manual is a dynamic document and is revised on a regular basis. Suggestions for additions and changes to this document are welcome and can be sent to the OBIS Capacity Development Task Team by email to training@obis.org or added as issues at https://github.com/iobis/manual/issues.

1.1 Guidelines on the sharing and use of data in OBIS

It is important that our data providers as well as all the data users are aware and agree on the OBIS guidelines on the sharing and use of data in OBIS, which was adoped at the 4th OBIS Steering Group.

1.2 Acknowledgements

This manual received contributions from: Leen Vandepitte, Mary Kennedy, Philip Goldstein, Pieter Provoost, Samuel Bosch, Ward Appeltans, Abby Benson, Yi-Ming Gan, Carolina Peralta Brichtova, Saara Suominen, Serita van der Wal, and Elizabeth Lawrence.

1.3 Data Policy

1.3.1 Guidelines on the sharing and use of data in OBIS

Adopted at SG-OBIS-IV (Feb 2015) and IODE-XXIII (March 2015).

The OBIS data policy is based on the principles of timely, free and unrestricted access to biodiversity data for the benefit of science and society, as defined in the:

- IOC data exchange policy
- IOC guidelines on transfer of marine technology
- IODE objectives
- OBIS vision and mission

Unless data are collected through activities funded by IOC/IODE, neither UNESCO, IOC, IODE, the OBIS Secretariat, nor its employees or contractors, own the data in OBIS and they take no responsibility for the quality of data or products based on OBIS, or the use or misuse that people may make of them nor can it

control or limit the use of any data or products accessible through its website, other than through the use of a published Data Sharing and Use Terms and Conditions.

1.3.1.1 Data sharing agreement

The data providers retain all rights and responsibilities associated with the data they make available to OBIS via the OBIS nodes. The OBIS nodes warrant that they have made the necessary agreements with the original data providers that it can make the data available to OBIS data under the following Creative Commons licenses:

- CC-0 (most preferred)
- CC-BY
- CC-BY-NC (least preferred)

The data providers are responsible for the completeness of the data and metadata profiles. When data is made available to OBIS, OBIS is granted permission to:

- Distribute the data via its data and information portal
- Build an integrated database, use the data for data quality control purposes, complement the data with other data such as climate variables and build value-added information products and services for science and decision-making
- Serve the data to other similar open-access networks such as GBIF in compliance with the terms and conditions for use set by the data providers.

In pursuance of copyright compliance, OBIS endeavours to secure permission from rights holders to ingest their datasets. In the event that the inclusion of a dataset in OBIS is challenged on the basis of copyright infringement, OBIS will follow a take-down policy until there is resolution.

1.3.1.2 Data use agreement

The data in OBIS are freely available to everyone, following the principles of equitable access and benefit sharing and supporting capacity development and participation of all IOC Member States in global programmes. However, data users are expected to give attribution to the data providers (see Citations) and the use of data from OBIS should happen in the light of fair use, i.e.:

- Recognize that the OBIS portal holds the master copy of the integrated database and hence users should refrain from online redistribution of the OBIS database. Because the OBIS database is updated regularly (every so months) with new datasets and revisions of existing datasets, copies of the OBIS database will become out of date quickly. If you wish to build access web services on top of OBIS, please contact the OBIS secretariat.
- Respect the data providers, and provide helpful feedback on data quality.
- In the case you are a custodian of biogeographic data yourself you should take action to also publish these data through OBIS.
- Consider sponsoring or partnering with OBIS and its OBIS nodes in grant proposal writing. Creating a global database like OBIS cannot happen without the, often voluntary, contribution of many scientists and data managers all over the world. Several activities, such as the coordination, data aggregation, quality control, database and website maintenance require resources including manpower at national and international level. A list of sponsors can be found here

For guidelines on how to cite data obtained from OBIS, see the Citing section of the Manual.

1.3.1.3 Disclaimer

Appropriate caution is necessary in the interpretation of results derived from OBIS. Users must recognize that the analysis and interpretation of data require background knowledge and expertise about marine biodiversity (including ecosystems and taxonomy). Users should be aware of possible errors, including in the use of species names, geo-referencing, data handling, and mapping. They should crosscheck their results for possible errors, and qualify their interpretation of any results accordingly.

Unless data are collected through activities funded by IOC/IODE, neither UNESCO, IOC, IODE, the OBIS Secretariat, nor its employees or contractors, own the data in OBIS and they take no responsibility for the quality of data or products based on OBIS, or the use or misuse.

1.4 Getting Help in OBIS

If you require additional assistance with OBIS we recommend you first get in touch with the most relevant OBIS node. We also have a **support channel** on Slack where you can communicate with the OBIS community for help. Please feel comfortable posting to this channel before reaching out to the OBIS Secretariat (helpdesk@obis.org). The OBIS community is quite active on Slack and GithHub (see below) so you are more likely to receive a quick answer to your question by posting in either place, as the Secretariat receives many requests.

You can submit issues and questions on relevant Github repositories:

- OBIS Manual
- OBIS issues GitHub repo
- OBIS quality control issues
- All other OBIS repositories

We strongly recommend creating a GitHub account to engage with the OBIS community, document issues, ask questions, find datasets that need endorsing, etc. GitHub gives threads a more permanent home and allows for open communication and transparency. If you are unfamiliar with GitHub, the Carpentries have these training resources which you can reference.

1.5 Frequently Asked Questions

1.5.0.1 General

- I have data and want to publish to OBIS - what do I do?
- Why is it important to share and format data?
- How do I handle sensitive data?
- Where can I make suggestions for improvements on this Manual?
- Where can I find OBIS related training videos?
- What are the responsibilities of node managers?
- Where can I find marine datasets linked to the OBIS network by the GBIF registry, that now require endorising?

1.5.0.2 Darwin Core

- Where can I learn about “Darwin Core”?
- I am having trouble understanding how Core and Extension tables relate to each other
- How does the OBIS format avoid redundancy in data
- How are extension tables (e.g. eMOF, occurrence) linked with the core table?
- What is the difference between Occurence Core and Event Core?

1.5.0.3 Formatting Data

Is there a checklist of all required Darwin Core fields for OBIS?

How does data flow in OBIS?

What should I do if I do not have the data for required fields by OBIS?

How do I construct an eventID?

How do I construct occurrenceID?

What data goes into Occurrence core (or extension) and how do I set up this file?

How do I set up an Event core table?

Do I have to provide decimalLatitude and decimalLongitude for the Event and Occurrence tables?

The answer may depend on your dataset structure, but generally, no. If you have Event core, then you do not need to repeat location information in the Occurrence table (but you can if you'd like). If you are using Occurrence core, then location information must be provided in the Occurrence table.

What data goes into extendedMeasurementOrFact and how do I set it up?

How do I format dates?

How do I handle historical data?

How do I convert coordinates to decimal degrees?

How do I convert different geographical formats to WGS84?

How do I compile acoustic, imaging, or other multimedia data for OBIS?

How do I compile habitat data for OBIS?

How do I compile tracking data for OBIS?

How do I compile DNA and genetic data for OBIS?

How do I document occurrences from unknown species, those new to science, or those with temporary names? e.g. Eurythenes sp. DISCOLL.PAP.JC165.674

Occurrences unknown or new to science should be documented according to recommendations by Horton et al. 2021. You should populate the `scientificName` field with the genus, and in `identificationQualifer` provide the ON sign 'sp.'. However you must also indicate the reason why species-level identification is unavailable. To do this, supplement 'sp.' with either stet. (stetit) or indet. (indeterminabilis). If neither of these are applicable, (e.g. for undescribed new species), add a unique taxon identifier code after 'sp.' to `identificationQualifer`. For example Eurythenes sp. DISCOLL.PAP.JC165.674.

Please avoid simple alphanumeric codes (i.e. Eurythenes sp. 1, Eurythenes sp. A). Similar to creating `eventIDs` or `occurrenceIDs`, you should strive to provide more complex and globally unique identifier. Identifiers could be constructed by combining higher taxonomic information with information related to a collection, institution, museum or collection code, sample number or museum accession number, expedition, dive number, or timestamp. This ensures namestrings will remain unique within a larger repositories like OBIS. It is also recommended to include these temporary names on specimen labels for physical specimens.

1.5.0.4 Vocabulary

How do I map Measurement or Fact terms in OBIS with preferred BODC vocabulary?

I can't find a suitable vocabulary, what do I do? How do I request a new vocabulary term?

Should I use taxon-specific P01 codes to populate for measurementTypeID? e.g. http://vocab.nerc.ac.uk/collection/P01/current/A15985A1

No. You should never use taxon-specific P01 codes. This is because the taxa are already identified in the Occurrence table, in the fields `scientificName` and `scientificNameID`.

How should I match raw data fields with Darwin Core terminology?

1.5.0.5 Tools

How do I use the WoRMS taxon match tool?

Can I fetch a full classification for a list of species from WoRMS?

What do I do if my scientificName does not return a match from WoRMS?

Where can I find DNA sequences published in OBIS?

Is there a template generator I can use to help create my Event, Occurrence, and eMoF tables?

Yes. There is an Excel template generator developed by Luke Marsden & Olaf Schneider as part of the Nansen Legacy project. Note this template generator is aimed at GBIF users, so make to account for and include required OBIS terms.

There is also this Excel to Darwin Core macro tool developed by GBIF Norway you can use to help generate templates.

How do I georeference locations, including text-based descriptions?

1.5.0.6 Quality Control

- How do I do data quality control?
- What are the OBIS quality control flags?
- Why are certain records dropped in OBIS?
- What do I do when I am uncertain about the:
  - Temporal range of a dataset OR eventDate
  - Geospatial location
  - Taxonomic identification
  - Individual count
- What do I do with freshwater species that are part of my marine dataset?

1.5.0.7 Publishing

- How do I add my data to the OBIS database?
- What metadata do I have to provide? Where? How?
- How do you know which license to choose?
- How do I access the IPT?
- How do I use the IPT?
- Are there instructions for IPT administrators?
- How do I add DOI to my dataset?
- How do I publish to both GBIF and OBIS?
- How do I update my already published dataset?

1.5.0.8 Accessing data in OBIS

How do I download data from OBIS?

How do I load the full (.csv) export of OBIS data?

Loading the entire OBIS dataset uses *a lot* of memory and is probably not feasible on most desktop computers. You have a few potential options depending on the use case: i) process the data in smaller batches, or ii) load the dataset into a local database such as SQLite and use SQL queries to analyze the data

Otherwise, we recommend you use the parquet download which is available here, instead of the CSV. Then in R, you can use the `arrow` package to work with parquet files. We also have a short tutorial on working with parquet files in R here, with an example application of this approach here (see first code block).

How can I use R to access OBIS data?

How do I use the OBIS API to fetch and filter data?

How do I contact the data provider?

How can I cite OBIS datasets and downloads?

What are the definitions of the field names in the downloads generated by OBIS?

How do I obtain a taxon checklist for an area?

There are a few possible ways to obtain a taxon checklist for a given area. We will obtain a checklist of species in the Albain EEZ as an example. To do this we will create a bounding box around our area of interest, and then apply filters to simplify the geometry.

```
library(mregions)
library(dplyr)
library(robis)
library(sf)
#obtain Albanian EEZ as sf
geom <- mr_shp(key = "MarineRegions:eez", filter = "Albanian Exclusive Economic Zone", maxFeatures = N
#get WKT for the bounding box
wkt <- st_as_text(st_as_sfc(st_bbox(geom)), digits = 6)
#fetch occurrences for bounding box
occ <- occurrence(geometry = wkt) %>%
  st_as_sf(coords = c("decimalLongitude", "decimalLatitude"), crs = 4326)
#filter using geometry
occ_filtered <- occ %>%
  filter(st_intersects(geometry, geom, sparse = FALSE)) %>%
  as_tibble() %>%
  select(-geometry)
#get taxa
alb_taxa <- occ_filtered %>%
  group_by(phylum, class, order, family, genus, species, scientificName) %>%
  summarize(records = n())
```

The dates look unusual in the download file. What are these, how do I convert them, and/or how do I obtain separate elements from them (e.g. month)?

The values in `date_start`, `date_mid`, and `date_end` are unix timestamps which have been calculated from the ISO date in the `eventDate` column. We can convert these numerical values to dates using the formula below.

```
=(E2/86400000)+DATE(1970,1,1)
```

If, when you apply this formula, you still see numbers, you will need to set the cell formatting to Date. Once you have dates, you can obtain, e.g. months for seasonal analyses using:

```
=MONTH(H2)
```

You can also use this tool to convert timestamps.

How do I filter by or obtain trait information for OBIS data (e.g. all benthic organisms)?

Currently, it is not possible to filter OBIS data by trait. To do this, we recommend using the traits database of the World Register of Marine Species. For example, searching by "functional group", you can specify benthos, plankton, nekton, etc.

Contributing data to OBIS

Chapter 2

What can you contribute and how?

Since 2000, OBIS has accepted, curated and published marine biodiversity data obtained by varied sources and methods. There is a common misconception that OBIS only accepts species occurrence data - however this is not true! OBIS can accept many types of marine data including:

- Presence/Absence
- Abundance, individual count
- Biomass
- Abiotic measurements
- Biotic measurements
- Sampling methods
- Sample processing methods
- Genetic data including sequences
- Data originating from historical records
- Tracking data
- Habitat data
- Acoustic data
- Imaging data
- Metadata describing the dataset and any project or programme related metadata

So if you have any of these types of marine data linked to your occurrence data and also want to contribute to OBIS - great! OBIS accepts data from any organization, consortium, project or individual who wants to contribute data. OBIS Data Sources are the authors, editors, and/or organisations that have published one or more datasets through OBIS. They remain the owners or custodians of the data, not OBIS!

OBIS harvests and publishes data from recognized IPTs from OBIS nodes or GBIF publishers. If you own data or have the right to publish data in OBIS, you can contact the OBIS secretariat or one of the OBIS nodes, or additionally a GBIF publisher. Your organization or programme can also become an OBIS node. An OBIS node usually publishes data from multiple data holders, effectively being a node in a network of data providers. So you may have to first find a relevant node before you get your data ready to publish.

To publish a dataset to OBIS, there are **five** main steps you must go through.

1. First, you must identify which OBIS node is best suited to host your published data. If you would like to publish to GBIF at the same time, that is also possible. If your organization is already affiliated with a GBIF node with which you must publish from, OBIS can also harvest from GBIF nodes.
2. Second, you must determine the structure of your data and which format will best suit your dataset. OBIS follows Darwin Core Archive (DwC-A) standards for datasets, and currently follows a star schema format. This format is based on relational databases. If you are unfamiliar with such database structures, or would like to refamiliarize yourself with them, please read here

3. Then, you need to actually format your data according to OBIS and DwC-A standards and guidelines
4. Once formatted, you should run a series of quality control measures to ensure you are not missing any required information and that all standards are being met. This helps ensure all data published in OBIS is formatted in a standardized way. When published in OBIS, OBIS provides a quality report to inform data owners and users of any quality control issues. By completing quality control before you publish your dataset you ensure there are fewer errors to fix later.
5. Now that your dataset is ready for publishing, the relevant metadata must be filled in, and then published on the previously identified IPT.

Each of these steps are covered in detail in the relevant sections of the manual. For an overview of this process see data management flow in OBIS.

2.1 Why publish data to OBIS

It is important to publish and ensure your dataset follows a universal standard for several reasons. The FAIR guiding principles for scientific data management and stewardship provide a good framework to understand the reasoning behind publishing data. FAIR stands for Findable, Accessible, Interoperable, and Reusable. Let's understand each aspect within the FAIR framework and how it is linked to publishing data in OBIS.

- **F - Findable**

Even if you publish your dataset on its own, publishing your data with OBIS will make your data more Findable (and Accessible) to a wider audience you might not have otherwise reached. By publishing your dataset to OBIS you are adding to a global database where your data can be found and analyzed alongside thousands of other datasets. For example, a dataset on marine invasive species in Venezuela was published July 20, 2022 and as of October 5, 2022 records of this dataset were included in 1,873 data download requests. This can save you time rather than handling individual data requests.

- **A - Accessible**

Similar to being Findable, OBIS makes your datasets more Accessible. Each dataset is given an identifier when you upload it on an IPT. Thus when users obtain data from OBIS, the original dataset can easily be identified and accessed. Data from OBIS is accessible in numerous ways, giving data users multiple avenues to potentially access your data.

- **I - Interoperable**

Using a standardized data format with controlled vocabularies will ensure your data are more Interoperable - more easily interpreted and processed by computers and humans alike. Increasingly, scientists use computer programs to conduct e-Science and collect data with algorithms. Formatting your data for OBIS will ensure it can be read and accessed by such programs as well as understood by users.

- **R - Reusable**

Publishing your data allows it to be Reused according to your chosen data usage license. Very likely you expended resources to collect your data and it would be a waste of those resources to leave your unique data unpublished and inaccessible for current and future generations. Likewise, it is better to preserve any data processing done to ensure your dataset is reproducible and/or verifiable. Finally, data in OBIS is often used in several assessment processes and used as information to support policy makers around the globe making informed decisions.

There are many other benefits of publishing in OBIS, even if you haven't published any work on it yet. This includes:

- Your dataset can be associated with a DOI, allowing for your dataset to be more easily cited. By ensuring your dataset citation is complete you will ensure you are being cited properly.
- Publishing your dataset with OBIS makes it easier to set it up as a Data paper, which generates value for you and other researchers.

- There are social benefits to data publishing as your work becomes integrated into a wider dataset. It gives both you and your data more visibility. This can lead to more opportunities for collaboration and further career development as a researcher or professional.
- Your data can be incorporated into larger analyses to better understand global ocean biodiversity, helping to shape regional and international policies.

2.2 How to handle sensitive data

We recognize that sometimes your dataset may contain sensitive information (e.g., location data on endangered or poached species), or perhaps your organization does not want certain details publicly accessible. Types of sensitive data include:

- Location data on endangered or protected species
- Information regarding a commonly poached species
- Species or locations that have an economic impact (positive or negative)

To accommodate sensitivity but still be able to contribute to OBIS, we suggest:

- Generalizing location information by: Obtaining regional coordinates using MarineRegions, Getty Thesaurus of Geographic Names, or Google Maps
- Using the OBIS Map tool to generate a polygon area with a Well-Known Text (WKT) representation of the geometry to paste into the `footprintWKT` field.
- Delay timing of publication (e.g., to accommodate mobile species)
- Submit your dataset, but mark it as private in the IPT so it is not published right away (i.e., until you set it as public). Alternatively, you can set a password on your dataset in order to share with specific individuals. Note that setting passwords will require some coordination with the IPT manager. By submitting your data to an IPT but not immediately publishing it, you can ensure that the dataset will be in a place to be incorporated at a later date when it is ready to be made public. This not only saves time and helps retain details while relatively fresh in your mind, but also ensures the dataset is still ready to be mobilized in case jobs are changed at a later date.

GBIF has created the following Best Practices for Generalizing Sensitive data which can provide you with additional guidance. `Chapman AD (2020) Current Best Practices for Generalizing Sensitive Species Occurrence Data. Copenhagen: GBIF Secretariat. https://doi.org/10.15468/doc-5jp4-5g10.`

2.3 OBIS Data Life Cycle

The basic data life cycle for contributions to OBIS can be broken down into six step-by-step phases:

1. Data structure
2. Data formatting
3. Quality control
4. Publishing
5. Data access (downloading)
6. Data visualization

Each of these phases are outlined in this manual and are composed of a number of steps which are covered in the relevant sections.

After you have decided on your data structure and have moved to the Data Formatting stage, you must first match the taxa in your dataset to a registered list. In formatting your dataset you will ensure the required OBIS terms and identifiers are mapped correctly to your data fields and records.

Depending on your data structure, you will then format data into a DwC-A format with the appropriate Core table (Event or Occurrence) with any applicable extension tables. Any biotic or abiotic measurements will

be moved into the extendedMeasurementOrFact table. Before proceeding to the publishing stage, there are a number of quality control steps to complete.

Once your data has been published, you and others can access datasets through various avenues and it becomes part of OBIS' global database!

This may seem like a daunting process at first glance, but this manual will walk you through each step, and the OBIS community is full of helpful resources. Throughout the manual you will find tutorials and tools to guide you from start to finish through the OBIS data life cycle.

2.3.0.1 Who is responsible for each phase?

Phases 1 through 3 are the responsibilities of the data provider, while Phases 3 and 4 are shared between the data provider and the node manager. Data users are involved in Phases 5 and 6.

The OBIS Secretariat is responsible for data processing and harvesting published resources.

2.4 Biodiversity data standards

From the very beginning, OBIS has championed the use of international standards for biogeographic data. Without agreement on the application of standards and protocols, OBIS would not have been able to build a large central database. OBIS uses the following standards:

- Darwin Core
- Ecological Metadata Language
- Darwin Core Archive and dataset structure

The following pages of this manual review each of these in turn. We show you how to apply these standards to format your data in the Data Formatting section.

We also provide some dataset examples for your reference.

2.4.1 Darwin Core

Contents

2.4.1.1 Introduction to Darwin Core

Darwin Core is a body of standards (i.e., identifiers, labels, definitions) that facilitate sharing biodiversity informatics. It provides stable terms and vocabularies related to biological objects/data and their collection.

Darwin Core is maintained by TDWG (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases). Stable terms and vocabularies are important for ensuring the datasets in OBIS have consistently interpretable fields. By following Darwin Core standards, both data providers and users can be certain of the definition and quality of data.

2.4.1.1.1 History of Darwin Core and OBIS The old OBIS schema was an OBIS extension to Darwin Core 1.2., which was based on Simple Darwin Core, a subset of Darwin Core which does not allow any structure beyond rows and columns. This old schema added some terms which were important for OBIS, but were not supported by Darwin Core at the time (e.g., start and end date and start and end latitude and longitude, depth range, lifestage, and terms for abundance, biomass and sample size).

In 2009, the Executive Committee of TDWG announced their ratification of an updated version of Darwin Core as a TDWG Standard. Ratified Darwin Core unifies specializations and innovations emerging from diverse communities, and provides guidelines for ongoing enhancement. The Darwin Core Quick Reference Guide links to TDWG's term definitions and related practices for Ratified Darwin Core. We will discuss the relevance of terms in this guide further below.

In December 2013, the 3rd session of the IODE Steering Group for OBIS agreed to transition OBIS globally to the TDWG-Ratified version of Darwin Core, and the mapping of the (old) OBIS specific terms to Darwin Core can be found here.

2.4.1.2 Darwin Core (DwC) terms

DwC terms correspond to the column names of your dataset and can be grouped according to class type for convenience, e.g., Taxa, Occurrence, Record, Location, etc. It is important to use DwC field names because only columns using Darwin Core terms as headers will be recognized.

A list of all possible Darwin Core terms can be found on TDWG. However, OBIS does not parse all terms (note this doesn't mean you cannot include them, they just will not be parsed when you publish to OBIS). Below is an overview of the most relevant Darwin Core terms to consider when contributing to OBIS, with guidelines regarding their use. We have also compiled a convenient checklist of OBIS-accepted terms, their DwC class type, and which OBIS file (Event Core, Occurrence, eMoF, etc.) it is likely to be found in.

Note that OBIS currently has seven required and one strongly recommended DwC term: `occurrenceID`, `eventDate`, `decimalLongitude`, `decimalLatitude`, `scientificName`, `occurrenceStatus`, `basisOfRecord`, `scientificNameID` (strongly recommended).

The following DwC terms are related to the Class *Taxon*:

- scientificName
- scientificNameID
- scientificNameAuthorship
- kingdom
- taxonRank
- taxonRemarks

The following DwC terms are related to the Class *Identification*:

- identifiedBy
- dateIdentified
- identificationReferences
- identificationRemarks
- identificationQualifier
- typeStatus

The following DwC terms are related to the Class *Occurrence*:

- occurrenceID

- occurrenceStatus
- recordedBy
- individualCount (OBIS recommends to add measurements to eMoF)
- organismQuantity (OBIS recommends to add measurements to eMoF)
- organismQuantityType (OBIS recommends to add measurements to eMoF)
- sex (OBIS recommends to add measurements to eMoF)
- lifeStage (OBIS recommends to add measurements to eMoF)
- behavior
- associatedTaxa
- occurrenceRemarks
- associatedMedia
- associatedReferences
- associatedSequences
- catalogNumber
- preparations

The following DwC terms are related to the Class *Record level*:

- basisOfRecord
- institutionCode
- collectionCode
- collectionID
- bibliographicCitation
- modified
- dataGeneralizations

The following DwC terms are related to the Class *Location*:

- decimalLatitude
- decimalLongitude
- coordinateUncertaintyInMeters
- geodeticDatum
- footprintWKT
- minimumDepthInMeters
- maximumDepthInMeters
- minimumDistanceAboveSurfaceInMeters
- maximumDistanceAboveSurfaceInMeters
- locality
- waterBody
- islandGroup
- island
- country
- locationAccordingTo
- locationRemarks
- locationID

The following DwC terms are related to the Class *Event*:

- parentEventID
- eventID
- eventDate
- type
- habitat
- samplingProtocol (OBIS recommends to add sampling facts to eMoF)
- sampleSizeValue (OBIS recommends to add sampling facts to eMoF)
- SampleSizeUnit (OBIS recommends to add sampling facts to eMoF)

- samplingEffort (OBIS recommends to add sampling facts to eMoF)

The following DwC terms are related to the Class *MaterialSample*:

- materialSampleID

2.4.1.3 Darwin Core guidelines

2.4.1.3.1 Taxonomy and identification `scientificName` (required term) should always contain the originally recorded scientific name, even if the name is currently a synonym. This is necessary to be able to track back records to the original dataset. The name should be at the lowest possible taxonomic rank, preferably at species level or lower, but higher ranks, such as genus, family, order, class etc. are also acceptable. We recommend to not include authorship in `scientificName`, and only use `scientificNameAuthorship` for that purpose. The `scientificName` term should only contain the name and not identification qualifications (such as ?, confer or affinity), which should instead be supplied in the `IdentificationQualifier` term, see examples below. `taxonRemarks` can capture comments or notes about the taxon or name.

A WoRMS LSID should be added in `scientificNameID` (strongly recommended term), OBIS will use this identifier to pull the taxonomic information from the World Register of Marine Species (WoRMS) into OBIS and attach it to your dataset. This information includes:

- Taxonomic classification (kingdom through species)
- The accepted name in case of invalid names or synonyms
- AphiaID
- IUCN red list category

LSIDs are persistent, location-independent, resource identifiers for uniquely naming biologically significant resources. More information on LSIDs can be found at www.lsid.info. For example, the WoRMS LSID for *Solea solea* is: urn:lsid:marinespecies.org:taxname:127160, and can be found at the bottom of each WoRMS taxon page, e.g. *Solea solea*.

`kingdom` and `taxonRank` can help us in identifying the provided `scientificName` in case the name is not available in WoRMS. `kingdom` in particular can help us find alternative genus-species combinations and avoids linking the name to homonyms. Please contact the WoRMS data management team (info@marinespecies.org) in case the scientificName is missing in WoRMS. `kingdom` and `taxonRank` are not necessary when a correct `scientificNameID` is provided.

OBIS recommends providing information about how an identification was made, for example by which ID key, species guide or expert; and by which method (e.g morphology vs. genomics), etc. The person's name who made the taxonomic identification can go in `identifiedBy` and *when* in `dateIdentified`. Use the ISO 8601:2004(E) standard for date and time, for instructions see Time. A list of references, such as field guides used for the identification can be listed in `identificationReferences`. Any other information, such as identification methods, can be added to `identificationRemarks`.

Examples:

scientificNameID	scientificName	kingdom	phylum	class
urn:lsid:marinespecies.org:taxname:142004	Yoldiella nana	Animalia	Mollusca	Bivalvia
urn:lsid:marinespecies.org:taxname:140584	Ennucula tenuis	Animalia	Mollusca	Bivalvia
urn:lsid:marinespecies.org:taxname:131573	Terebellides stroemii	Animalia	Annelida	Polychaeta

order	family	genus	specificEpithet	scientificNameAuthorship
Nuculanoida	Yoldiidae	Yoldiella	nana	(Sars M., 1865)
Nuculoida	Nuculidae	Ennucula	tenuis	(Montagu, 1808)
Terebellida	Trichobranchidae	Terebellides	stroemii	Sars, 1835

Data from Benthic fauna around Franz Josef Land.

If the record represents a nomenclatural type specimen, the term `typeStatus` can be used, e.g. for holotype, syntype, etc.

In case of low confidence identifications, and the scientific name contains qualifiers such as *cf.*, *?* or *aff.*, then this name should go in `identificationQualifier`, and `scientificName` should contain the name of the lowest possible taxon rank that refers to the most accurate identification. E.g. if the specimen was accurately identified down to genus level, but not species level, then the scientificName should contain the name of the genus, the scientificNameID should contain the LSID the genus and the `identificationQualifier` should contain the low confidence species name combined with *?* or other qualifiers. The table below shows a few examples:

The use and definitions for additional ON signs (`identificationQualifier`) can be found in Open Nomenclature in the biodiversity era, which provides examples for using the main Open Nomenclature qualifiers associated with *physical specimens*. The publication Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications provides examples and definitions for `identificationQualifiers` for *non-physical specimens (image-based)*.

Examples:

scientificName	scientificNameAuthorship	scientificNameID	taxonRank	identificationQualifier	taxonConceptID
Pelagia	Péron & Lesueur, 1810	urn:lsid:marinespecies.org:taxname:135262	genus	gen. nov.	Pelagia gen. nov.
Pelagia benovici	Piraino, Aglieri, Scorrano & Boero, 2014	urn:lsid:marinespecies.org:taxname:851656	species	sp. nov	Pelagia benovici sp. nov
Gadus	Linnaeus, 1758	urn:lsid:marinespecies.org:taxname:125732	genus	cf. morhua	Gadus cf. morhua
Polycera	Cuvier, 1816	urn:lsid:marinespecies.org:taxname:138369	genus	cf. hedgpethi	Polycera cf. hedgpethi
Tubifex	Lamarck, 1816	urn:lsid:marinespecies.org:taxname:137392	genus	?	Tubifex tubifex(Müller, 1774)?
Tubifex	Lamarck, 1816	urn:lsid:marinespecies.org:taxname:137392	genus	sp. inc.	Tubifex tubifex(Müller, 1774)sp. inc.
Brisinga	Asbjørnsen, 1856	urn:lsid:marinespecies.org:taxname:123210	genus	gen. inc.	Brisinga gen. inc.
Uroptychus compressus	Baba & Wicksten, 2019	urn:lsid:marinespecies.org:taxname:1332465	genus	sp. inc.	Uroptychus compressus sp. inc.
Eurythenes	S. I. Smith in Scudder, 1882	urn:lsid:marinespecies.org:taxname:101607	genus	sp. DISCOLL.PAP.JC165.674	Eurythenes sp.DISCOLL.PAP.JC165.674
Paroriza	Hérouard, 1902	urn:lsid:marinespecies.org:taxname:123467	genus	sp.[unique123]aff.pallens	Paroriza sp.[unique123]aff. pallens
Aristeidae	Wood-Mason in Wood-Mason & Alcock, 1891	urn:lsid:marinespecies.org:taxname:106725	family	stet.	Aristeidae stet.
Nematocarcinus	Milne-Edwards, 1881	urn:lsid:marinespecies.org:taxname:107015	genus	sp.indet.	Nematocarcinus sp.indet.
Brisinga	Asbjørnsen, 1856	urn:lsid:marinespecies.org:taxname:123210	genus	gen.inc.	Brisinga gen.inc.
Brisinga costata	Verrill, 1884	urn:lsid:marinespecies.org:taxname:17825	species	sp.inc.	Brisinga costata sp.inc.

2.4.1.3.2 Occurrence `occurrenceID` (required term) is an identifier for the occurrence record and should be persistent and globally unique. If the dataset does not yet contain (globally unique) occurrenceIDs, then they should be created. Guideline for ID creation can be found here

`occurrenceStatus` (required term) is a statement about the presence or absence of a taxon at a location. It is an important term, because it allows us to distinguish between presence and absence records. It is a required term and should be filled in with either `present` or `absent`.

A few terms related to quantity: `organismQuantity` and `organismQuantityType`, have been added to the TDWG ratified Darwin Core. This is a lot more versatile than the older `individualCount` field. However, OBIS recommends to use the Extended MeasurementorFact extension for quantitative measurements because of the standardization of terms and the fact that you can link these measurements to sampling events and factual sampling information.

Please take note that OBIS recommends all quantitative measurements and sampling facts to be placed in the `ExtendedMeasurementOrFact` extension and not in the Darwin Core files.

In the case specimens were collected and stored (e.g. museum collections), the `catalogNumber` and `preparations` terms can be used to provide the identifier for the record in the collection and to document the preparation and preservation methods. The term `typeStatus` see above (under identification) can be used in this context too.

Both `associatedMedia`, `associatedReferences` and `associatedSequences` are global unique identifiers or URIs pointing to respectively associated media (e.g. online image or video), associated literature (e.g. DOIs) or genetic sequence information (e.g. GenBANK ID).

`associatedTaxa` include a list (concatenated and separated) of identifiers or names of taxa and their associations with the Occurrence, e.g. the species occurrence was associated to the presence of kelp such as *Laminaria digitata*.

The recommended vocabulary for `sex` see BODC vocab : S10, for `lifeStage` see BODC vocab: S11, `behavior` (no vocab available), and `occurrenceRemarks` can hold any comments or notes about the Occurrence.

`recordedBy` can hold a list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original Occurrence. The primary collector or observer, especially one who applies a personal identifier (recordNumber), should be listed first.

Example:

collectionCode	occurrenceID	catalogNumber	occurrenceStatus
SluiceDock_benthic_1976/1981	SluiceDock_benthic_1976_1	SluiceDock_benthic_1976_1	present
SluiceDock_benthic_1976/1981	SluiceDock_benthic_1976_2	SluiceDock_benthic_1976_2	present
SluiceDock_benthic_1976/1981	SluiceDock_benthic_1979-07/1980-06_1	SluiceDock_benthic_1979-07/1980-06_1	present

Data from A summary of benthic studies in the sluice dock of Ostend during 1976-1981.

2.4.1.3.3 Record level terms `basisOfRecord` (required term) specifies the nature of the record, i.e. whether the occurrence record is based on a stored specimen or an observation. In case the specimen is collected and stored in a collection (e.g. at a museum, university, research institute), the options are:

- `PreservedSpecimen` e.g. preserved in ethanol, tissue etc.
- `FossilSpecimen` a fossil, which allows OBIS to make the distinction between the date of collection and the time period the specimen was assumed alive
- `LivingSpecimen` an intentionally kept/cultivated living specimen e.g. in an aquarium or culture collection.

In case no specimen is deposited, the basis of record is either `HumanObservation` (e.g bird sighting, benthic sample but specimens were discarded after counting), or `MachineObservation` (e.g. for occurrences based on automated sensors such as image recognition, etc). For records pertaining to genetic samples, basisOfRecord can be `MaterialSample` (e.g. in the DNA-derived data extension).

When the basisOfRecord is either a *preservedSpecimen*, *LivingSpecimen* or *FossilSpecimen* please also add the `institutionCode`, `collectionCode` and `catalogNumber`, which will enable people to visit the collection and re-examine the material. Sometimes, for example in case of living specimens, a dataset can contain records pointing to the origin, the in-situ sampling position as well as a record referring to the ex-situ collection. In this case please add the event type information in `eventRemarks` (see OBIS manual: event).

`institutionCode` identifies the custodian institute (often by acronym), `collectionCode` identifies the collection or dataset within that institute. Collections cannot belong to multiple institutes, so all records within a collection should have the same `institutionCode`. The `collectionID` is an identifier for the record within the dataset or collection.

`bibliographicCitation` allows for providing different citations on record level, while a single citation for the entire dataset can and should be provided in the metadata (see EML). The citation at record level can have the format of a chapter in a book, where the book is the dataset citation. The record citation will have preference over the dataset citation. We do not, however, recommend to create different citations for every record, as this will explode the number of citations and will hamper the re-use of data.

`modified` is the most recent date-time on which the resource was changed. It is required to use the ISO 8601:2004(E) standard, for instructions see Time.

`dataGeneralizations` refers to actions taken to make the shared data less specific or complete than in its original form. Suggests that alternative data of higher quality may be available on request. This can be the case for occurrences of vulnerable or endangered species and there positions are converted to the center of grid cells.

2.4.1.3.4 Location `decimalLatitude` and `decimalLongitude` (required terms) are the geographic latitude and longitude (in decimal degrees), using the spatial reference system given in `geodeticDatum` of the geographic center of a Location. The number of decimals should be appropriate for the level of uncertainty in `coordinateUncertaintyInMeters` (at least within an order of magnitude). `coordinateUncertaintyInMeters` is the radius of the smallest circle around the given position containing the whole location. Regarding `decimalLatitude`, positive values are north of the Equator, negative values are south of it. All values lie between -90 and 90, inclusive. Regarding `decimalLongitude`, positive values are east of the Greenwich Meridian, negative values are west of it. All values lie between -180 and 180, inclusive.

In OBIS, the spatial reference system to be documented in `geodeticDatum` is EPSG:4326. Coordinates in degrees/minutes/seconds can be converted to decimal degrees using our coordinates tool. We also provide a tool to check coordinates or to determine coordinates for a location (point, transect or polygon) on a map. This tool also allows geocoding location names using marineregions.org.

The name of the place or location can be provided in `locality`, and if possible linked by a `locationID` using a persistent ID from a gazetter, such as the MRGID from MarineRegions. If the species occurrence only contains the name of the `locality`, but not the exact coordinates, we recommend using a geocoding service to obtain the coordinates. Marine Regions has a search interface for geographic names, and provides coordinates and often precision in meters, which can go into `coordinateUncertaintyInMeters`. Another option is to use the Getty Thesaurus of Geographic Names or Google Maps: after looking up a location, the decimal coordinates can be found in the page URL. Additional information about the locality can also be stored in DwC terms such as `waterBody`, `islandGroup`, `island` and `country`. `locationAccordingTo` should provide the name of the gazetteer that is used to obtain the coordinates for the locality.

`locationID` is an identifier for the set of location information (e.g. station ID, or MRGID from marineregions), for example the Balearic Plain has MRGID: http://marineregions.org/mrgid/3956.

A Well-Known Text (WKT) representation of the shape of the location can be provided in `footprintWKT`. This is particularly useful for tracks, transects, tows, trawls, habitat extent or when an exact location is not known. WKT strings can be created using our WKT tool. This tool also calculates a midpoint and a radius, which can then be added to `decimalLongitude`, `decimalLatitude`, and `coordinateUncertaintyInMeters` respectively. There is also an R tool to calculate the centroid and radius for WKT polygons. wktmap.com can be used to visualize and share WKT strings.

Some examples of WKT strings:

```
LINESTRING (30 10, 10 30, 40 40)
POLYGON ((30 10, 40 40, 20 40, 10 20, 30 10))
MULTILINESTRING ((10 10, 20 20, 10 40),(40 40, 30 30, 40 20, 30 10))
MULTIPOLYGON (((30 20, 45 40, 10 40, 30 20)),((15 5, 40 10, 10 20, 5 10, 15 5)))
```

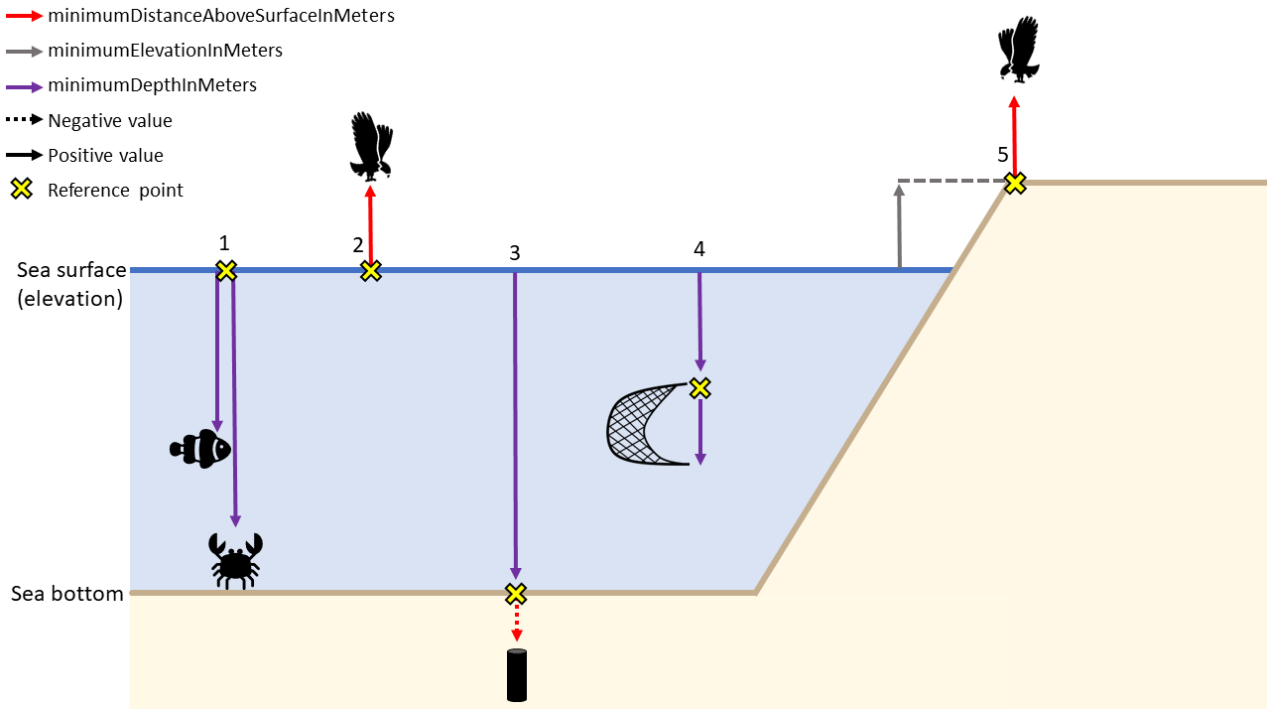
Example:

decimalLatitude	decimalLongitude	geodeticDatum	coordinateUncertaintyInMeters	footprintWKT	footprintSRS
38.698	20.95	EPSG:4326	75033.17	LINESTRING (20.31 39.15, 21.58 38.24)	EPSG:4326
42.72	15.228	EPSG:4326	154338.87	LINESTRING (16.64 41.80, 13.82 43.64)	EPSG:4326
39.292	20.364	EPSG:4326	162083.27	LINESTRING (19.05 40.34, 21.68 38.25)	EPSG:4326

Data from Adriatic and Ionian Sea mega-fauna monitoring employing ferry as platform of observation along the Ancona-Igoumenitsa-Patras lane, from December 2014 to December 2018.

Keep in mind while filling in `minimumDepthInMeters` and `maximumDepthInMeters` that this should be the depth at which the **sample was taken** and not the water column depth at that location. When fillling in any depth fields (`minimumDepthInMeters`, `maximumDepthInMeters`, `minimumDistanceAboveSurfaceInMeters`, and `maximumDistanceAboveSurfaceInMeters`), you should also consider which information is needed to fully understand the data. In most cases (e.g. scenario 1 and 4 in the figure below), providing `minimumDepthInMeters` and `maximumDepthInMeters` is sufficient for observations of organisms at particular depths. However, in cases where an occurrence is above the sea surface, e.g. flying birds (scenario 2 and 5), you should populate `minimumDistanceAboveSurfaceInMeters`, `maximumDistanceAboveSurfaceInMeters`, and, where relevant, you should also include `minimumElevationInMeters` and `maximumElevationInMeters`.

The `minimumDistanceAboveSurfaceInMeters` and `maximumDistanceAboveSurfaceInMeters` is the distance, in meters, above or below a reference surface or reference point. The reference surface is determined by the depth or elevation. If the depth and elevation are 0, then the reference surface is the sea surface. If a depth is given, the reference surface is the location of the depth. This can be especially useful for sediment cores taken from the sea bottom (scenario 3 in figure below). If no depth is given, then the elevation is the reference surface (scenario 5).



Depth scenario examples:

Scenario	minimumDepthInMeters	maximumDepthInMeters	minimumDistanceAboveSurfaceInMeters	maximumDistanceAboveSurfaceInMeters	minimumElevationInMeters	maximumElevationInMeters
1	40, 90	50, 100	-	-	0	0
2	0	0	10	15	0	0
3	100	100	0	-1.5	0	0
4	20	22	-	-	0	0
5	0	0	10	15	10	10

2.4.1.3.5 Event `eventID` is an identifier for the sampling or observation event. `parentEventID` is an identifier for a parent event, which is composed of one or more sub-sampling (child) events (eventIDs). See identifiers for details on how these terms can be constructed.

`habitat` is a category or description of the habitat in which the Event occurred (e.g. benthos, seamount, hydrothermal vent, seagrass, rocky shore, intertidal, ship wreck etc.)

2.4.1.3.6 Time The date and time at which an occurrence was recorded goes in `eventDate`. This term uses the ISO 8601 standard and OBIS recommends using the extended ISO 8601 format with hyphens.

More specific guidelines on formatting dates and times can be found in the Common Data formatting issues page

2.4.1.3.7 Sampling Information on `sampleSizeValue` and `sampleSizeUnit` is very important when an organism quantity is specified. However, with OBIS-ENV-DATA it was felt that the extended Measurementor-Fact (eMoF) extension would be better suited than the DwC Event Core to store the sampled area and/or volume because in some cases sampleSize by itself may not be detailed enough to allow interpretation of the sample. For instance, in the case of a plankton tow, the volume of water that passed through the net is relevant. In case of Niskin bottles, the volume of sieved water is more relevant than the actual volume in the bottle. In these examples, as well as generally when recording sampling effort for all protocols, eMoF enables greater flexibility to define parameters, as well as the ability to describe the entire sample and treatment protocol through multiple parameters. eMoF also allows you to standardize your terms to a controlled vocabulary.

The next chapter deals with the metadata (description of the dataset) in Ecological Metadata Language.

2.4.2 Darwin Core Archive

Contents

2.4.2.1 Darwin Core Archive

Darwin Core Archive (DwC-A) is the standard for packaging and publishing biodiversity data using Darwin Core terms. It is the preferred format for publishing data in OBIS and GBIF. The format is described in the Darwin Core text guide. A Darwin Core Archive contains a number of text files, including data tables formatted as CSV.

The conceptual data model of the Darwin Core Archive is a star schema with a single core table, for example containing occurrence records or event records, at the center of the star. Extension tables can optionally be associated with the core table. It is not possible to link extension tables to other extension tables (to form a so-called snowflake schema). There is a one-to-many relationship between the core and extension records, so each core record can have zero or more extension records linked to it, and each extension record must be linked to exactly one core record. Definitions for the core and extension tables can be found here.

Besides data tables, a Darwin Core Archive also contains two XML files: one file which describes the archive and data file structure (`meta.xml`), and one file which contains the dataset's metadata (`eml.xml`).

Figure: structure of a Darwin Core Archive.

2.4.2.2 OBIS holds more than just species occurrences: the ENV-DATA approach

Data collected as part of marine biological research often include measurements of habitat features (such as physical and chemical parameters of the environment), biotic and biometric measurements (such as body size, abundance, biomass), as wel as details regarding the nature of the sampling or observation methods, equipment, and sampling effort.

In the past, OBIS relied solely on the Occurrence Core, and additional measurements were added in a structured format (e.g. JSON) in the Darwin Core term `dynamicProperties` inside the occurrence records. This approach

had significant downsides: the format is difficult to construct and deconstruct, there is no standardization of terms, and attributes which are shared by multiple records (think sampling methodology) have to be repeated many times. The formatting problem can be addressed by moving measurements to a MeasurementOrFacts extension table, but that doesn't solve the redundancy and standardization problems.

With the release and adoption of a new core type Event Core it became possible to associate measurements with nested events (such as cruises, stations, and samples), but the restrictive star schema of Darwin Core archive prohibited associating measurements with the event records in the Event core as well as with the occurrence records in the Occurrence extension. For this reason an extended version of the existing MeasurementOrFact extension was created.

2.4.2.2.1 ExtendedMeasurementOrFact Extension (eMoF) As part of the IODE pilot project Expanding OBIS with environmental data OBIS-ENV-DATA, OBIS introduced a custom ExtendedMeasurementOrFact or eMoF extension, which extends the existing MeasurementOrFact extension with 4 new terms:

- `occurrenceID`
- `measurementTypeID`
- `measurementValueID`
- `measurementUnitID`

The `occurrenceID` term is used to circumvent the limitations of the star schema, and link measurement records in the ExtendedMeasurementOrFact extension to occurrence records in the Occurrence extension. Note that in order to comply with the Darwin Core Archive standard, these records still need to link to an event record in the Event core table as well. Thanks to this term we can now store a variety of measurements and facts linked to either events or occurrences:

- organism quantifications (e.g. counts, abundance, biomass, % live cover, etc.)
- species biometrics (e.g. body length, weight, etc.)
- facts documenting a specimen (e.g. living/dead, behaviour, invasiveness, etc.)
- abiotic measurements (e.g. temperature, salinity, oxygen, sediment grain size, habitat features)
- facts documenting the sampling activity (e.g. sampling device, sampled area, sampled volume, sieve mesh size).

Figure: Overview of an OBIS-ENV-DATA format. Sampling parameters, abiotic measurements, and occurrences are linked to events using the eventID (full lines). Biotic measurements are linked to occurrences using the new occurrenceID field of the ExtendedMeasurementOrFact Extension (dashed lines).

2.4.2.2.2 eDNA & DNA derived data Extension DNA derived data are increasingly being used to document taxon occurrences. To ensure these data are useful to the broadest possible community, GBIF published a guide entitled Publishing DNA-derived data through biodiversity data platforms. This guide is supported by the DNA derived data extension for Darwin Core, which incorporates MIxS terms into the Darwin Core standard. eDNA and DNA derived data is linked to occurrence data with the use of `occurrenceID` and/or `eventID`. Refer to the Examples: ENV-DATA and DNA derived data for use case examples of eDNA and DNA derived data.

2.4.2.2.3 A special case: habitat types Including information on habitats (biological community, biotope, or habitat type) is possible and encouraged with the use of Event Core. However, beware the unconstrained nature of the terms `measurementTypeID`, `measurementValueID`, and `measurementUnitID` which can lead to inconsistently documented habitat measurements within the Darwin Core Archive standard. To ensure this data is more easily discoverable, understood or usable, refer to Examples: habitat data and/or Duncan et al. (2021) for use case examples and more details.

2.4.2.2.4 Recommended reading

- De Pooter et al. 2017. Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. Biodiversity Data Journal 5: e10989. hdl.handle.net/10.3897/BDJ.5.e10989
- Duncan et al. (2021). A standard approach to structuring classified habitat data using the Darwin Core Extended Measurement or Fact Extension. EMODnet report. *(Note you must refine search to Technical Reports from 2021 to identify Duncan et al.'s report)*

2.4.3 Relational databases: the underlying framework of OBIS

If you are not familiar with relational databases, it can be difficult to understand the underlying framework OBIS relies on. This section will help you understand relational databases, how they relate to OBIS, the data you will format for OBIS, and the data you may download from OBIS.

Why do we use relational databases in the first place? You are probably familiar with flat databases which contain all data in one table - this is likely how your own data are formatted. Relational databases instead consist of multiple data tables that each contain *related* information. When all this information is presented in one table, the table becomes larger, very complicated, and the likelihood of data duplication increases. Relational databases seek to simplify complexities and reduce redundancy by allowing information to be self-contained, but linked to each other.

You can think of a relational database as separate Excel sheets or data tables that are related to each other. One data table could be a "core" table, whereas others are "extensions". Sometimes the relationships between core and extension tables are hierarchical, but this is not always the case. There is, however, always a *relationship* linking core and extension tables.

Let's review core and extension tables and how we use them for OBIS.

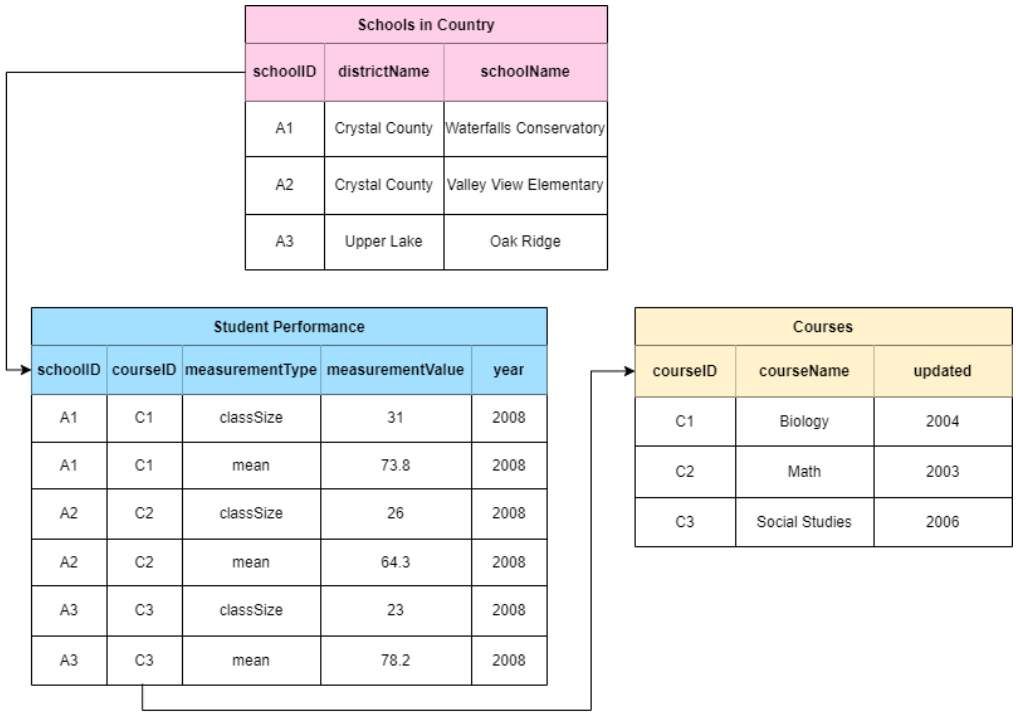
Core tables contain information that is applicable to **all** extension tables, and extension tables contain more information about the records within the Core table. Each table, whether core or extension, contains records and attributes. Each row is a record (e.g., a sampling event, a species' occurrence), whereas each column is an attribute (e.g., a date, a measurement).

Records between tables are linked to each other by the use of *identifiers*. A description of measurements pertaining to a record in an Extension table will have the same identifier as the record it is describing in the Core table. By using identifiers to link records, we reduce data repetition, see below for examples. In the Darwin Core format that OBIS uses, the core table is either Event or Occurrence, and datasets can have one, none, or more extension tables. Further explanation of data formatting in OBIS is covered in the Data Formatting section of the OBIS manual.

Let's review an example to fully understand how relational databases work. We will look at a simple relational database used by a fictional country that tracks student performance in three different courses between three schools. Rather than trying to contain information about each school, course, and student performance in one place, this information is split into three separate tables. We see that the pink table gives us information about each school - its name, and the district it belongs to. Each school also has a schoolID, an identifier linking to the blue table where we can see student performance (course mean) in each course, the class size, and year. You will notice that the course mean and class size are bundled under columns called measurementType and measurementValue. These are similar to the eMoF vocabularies and are integral to reducing repeated data, especially when one dataset has reoccurring information. Finally we see that the courseID in the blue table links to the yellow one with the courseID identifier, giving us information about each course.

A fourth table could easily be created to track total school population size through time. In contrast, if this information was only presented in the pink Schools in Country table, the school information would be duplicated as you add rows for each year. In this way, you can easily see how useful relational databases are. Of course, this is a simplified example but it demonstrates how related tables can be linked by identifiers to reduce table complexity and data replication.

We elaborate on how this structure is applied within OBIS here.



Schools in Country

schoolID	districtName	schoolName
A1	Crystal County	Waterfalls Conservatory
A2	Crystal County	Valley View Elementary
A3	Upper Lake	Oak Ridge

Student Performance

schoolID	courseID	measurementType	measurementValue	year
A1	C1	classSize	31	2008
A1	C1	mean	73.8	2008
A2	C2	classSize	26	2008
A2	C2	mean	64.3	2008
A3	C3	classSize	23	2008
A3	C3	mean	78.2	2008

Courses

courseID	courseName	updated
C1	Biology	2004
C2	Math	2003
C3	Social Studies	2006

Figure 2.1: *An example of how a relational database works. Three tables show the (1) student performance (blue table) in (2) different schools (pink table) in a fictional country, and (3) the names of the courses (yellow table). Information between each table is linked by the use of identifiers, indicated by the arrows*

Note that when OBIS harvests data, datasets are flattened - i.e., all separate data tables are combined into one. This is the kind of file you will receive when you download data from OBIS. The reason for this is that querying relational databases significantly reduces computational time, as opposed to querying a flat database. Relational databases also facilitate requests for subsets that meet particular criteria - e.g., all data from Norway for one species above a certain depth.

2.4.3.1 How to avoid redundancy

Avoiding redundancy and data duplication within your dataset is built into the OBIS data structure. Utilizing the ENV-DATA approach, which delineates relationships between the core table and extension tables, we can limit the repetition of data.

For example, let us consider the dates of a ship cruise where a series of bottom trawls were taken. The sampling information (e.g., date range, equipment used, etc.) for each species collected in these trawls is the same. Because of this, we know we are dealing with unique sampling events and thus we will use Event core. So, our Event core table will contain all information related to the sampling events (e.g., date, location). Then, information pertaining to each collected species (e.g., abundance, biomass, sampling methods, etc.) will be placed in an extension, the (Extended)MeasurementOrFact table. Here, each measurement for each species and sample will occur on a separate record. These records will be linked to the correct sampling event in the Event core by an identifier - the eventID. If we were to put this data in one file, the fields related to date and location (e.g., eventDate, decimalLongitude, decimalLatitude, etc.) would be repeated for each species.

Let's consider another example. If you took one temperature measurement from the water column where you took your sample, each species found in that sample would have the **same** temperature measurement. By linking such measurements to the *event* instead of each *occurrence*, we are able to reduce the amount of data being repeated.

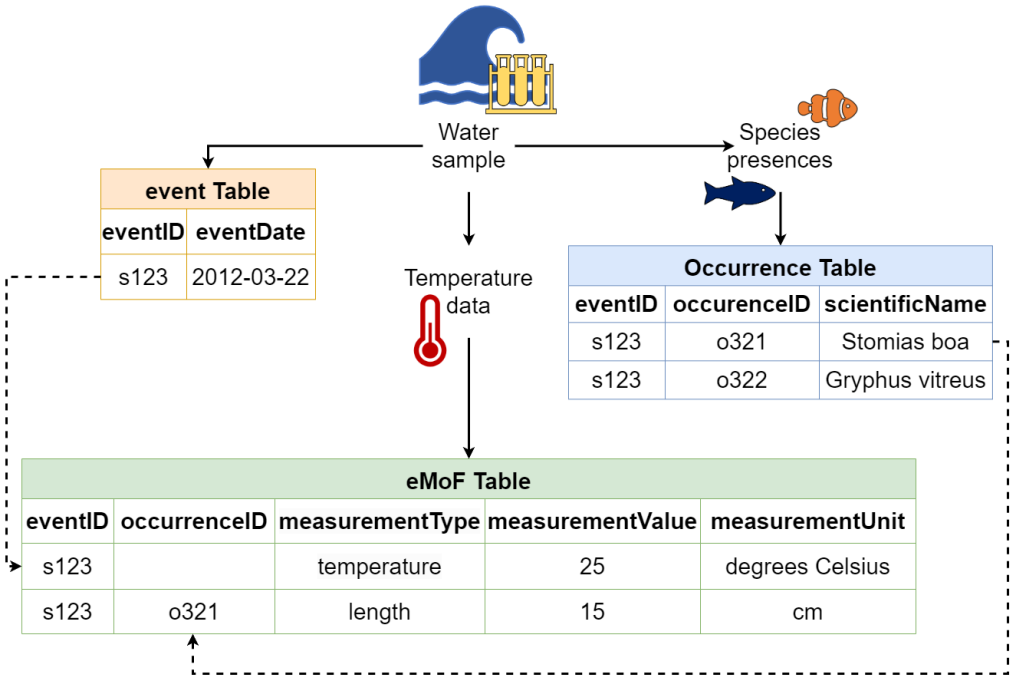


Figure 2.2: *Example of how the sample data is distributed to Core and Extension tables, and how these tables are connected in OBIS*

An advantage of structuring data this way is that if any mistakes are made, you only need to correct it once! So you can see that using relational event structures (when applicable) in combination with extension files can really simplify and reduce the number of times data are repeated.

Caveat: However we would like to note that in some cases, data duplication may occur due to the star schema structure. For example, when publishing DNA-derived data, Occurrence core will have to be used, which necessitates the repetition of event data for each occurrence record.

2.4.4 Ecological Metadata Language

OBIS (and GBIF) uses the Ecological Metadata Language (EML) as its metadata standard, which is specifically developed for the earth, environmental and ecological sciences. It is based on prior work done by the Ecological Society of America and associated efforts. EML is implemented as XML. See more information on EML.

OBIS uses the GBIF EML profile (version 1.1). In case data providers use ISO19115/ISO19139, there is a mapping available here.

For OBIS, the following 4 terms are the bare minimum required: `Title`, `Citation`, `Contact` and `Abstract`. Below is an overview of all the EML terms used to describe datasets:

- `title [xml:lang="..."]`: A good descriptive `title` is indispensable and can provide the user with valuable information, making the discovery of data easier. Multiple titles may be provided, particularly when trying to express the title in more than one language (use the "xml:lang" attribute to indicate the language if not English/en).
- `creator` ; `metadataProvider` ; `associatedParty` ; `contact` : These are the people and organizations responsible for the dataset resource, either as the creator, the metadata provider, contact person or any other association. The following details can be provided:
  - `individualName`

    * `givenName`
    * `surName`
  - `organizationName`: Name of the institution.
  - `positionName`: to be used as alternative to persons names (leave `individualName` blank and use `positionName` instead e.g. data manager).
  - `address`
    * `deliveryPoint`
    * `city`
    * `administrativeArea`
    * `postalCode`
    * `country`
  - `phone`
  - `electronicMailAddress`
  - `onlineUrl` : personal website
  - `role`: used with `associatedParty` to indicate the role of the associated person or organization.
  - `userID`: e.g. ORCID.
    * `directory`

- `pubDate`: The date that the resource was published. Use ISO 8601.
- `language`: The language in which the resource (not the metadata document) is written. Use ISO language code.
- `abstract` : Brief description of the data resource.
  - `para`
- `keywordSet`
  - `keyword` : Note only one keyword per keyword field is allowed.
  - `keywordThesaurus` : e.g. ASFA
- `additionalInfo` : OBIS checks this EML field for harvesting. It should contain *marine, harvested by iOBIS*.
  - `para`
- `coverage`
  - `geographicCoverage`
    * `geographicDescription`: a short text description of the area. E.g. the river mounth of the Scheldt Estuary.
    * `boundingCoordinates`
      · `westBoundingCoordinate`
      · `eastBoundingCoordinate`
      · `northBoundingCoordinate`
      · `southBoundingCoordinate`
  - `temporalCoverage` : Use ISO 8601
    * `singleDateTime`
    * `rangeOfDates`
      · `beginDate`
        · `calendarDate`
      · `endDate`
        · `calendarDate`
  - `taxonomicCoverage`: taxonomic information about the dataset. It can include a species list.
    * `generalTaxonomicCoverage`
    * `taxonomicClassification`

            - taxonRankName
            - taxonRankValue
            - commonName

- intellectualRights: Statement about IPR, Copyright or various Property Rights. Also read the guidelines on the sharing and use of data in OBIS.
  - para
- purpose: A description of the purpose of this dataset.
  - para
- methods
  - methodStep: Descriptions of procedures, relevant literature, software, instrumentation, source data and any quality control measures taken.
  - sampling: Description of sampling procedures including the geographic, temporal and taxonomic coverage of the study.
  - studyExtent: Description of the specific sampling area, the sampling frequency (temporal boundaries, frequency of occurrence), and groups of living organisms sampled (taxonomic coverage).
  - samplingDescription: Description of sampling procedures, similar to the one found in the methods section of a journal article.
    * para
  - qualityControl: Description of actions taken to either control or assess the quality of data resulting from the associated method step.
- project
  - title
  - identifier
  - personnel: The personnel field is used to document people involved in a research project by providing contact information and their role in the project.
  - description
  - funding: The funding field is used to provide information about funding sources for the project such as: grant and contract numbers; names and addresses of funding sources.
    * para
  - studyAreaDescription
  - designDescription: The description of research design.
- maintenance
  - description
    * para
  - maintenanceUpdateFrequency
- additionalMetadata
  - metadata
    * dateStamp: The dateTime the metadata document was created or modified (ISO 8601).
    * metadataLanguage: The language in which the metadata document (as opposed to the resource being described by the metadata) is written
    * hierarchyLevel
      - citation : A single citation for use when citing the dataset. The IPT can also auto-generate a citation based on the metadata (people, title, organization, onlineURL, DOI etc).
      - bibliography: A list of citations that form a bibliography on literature related / used in the dataset
      - resourceLogoUrl: URL of the logo associated with a dataset.
      - parentCollectionIdentifier

        - `collectionIdentifier`
        - `formationPeriod`: Text description of the time period during which the collection was assembled. E.g., “Victorian”, or “1922 - 1932”, or “c. 1750”.
        - `livingTimePeriod`: Time period during which biological material was alive (for palaeontological collections).
        - `specimenPreservationMethod`
        - `physical`
          - `objectName`
          - `characterEncoding`
          - `dataFormat`
            - `externallyDefinedFormat`
            - `formatName`
          - `distribution`: URL links
            - `online`
              - `url function="download"`
              - `url function="information"`

- `alternateIdentifier`: It is a Universally Unique Identifier (UUID) for the EML document and not for the dataset. This term is optional.

2.4.4.1 Metadata Sections

There are several categories/pages for metadata you must provide, which includes basic information about the:

- Dataset and data provider
- Geographic/taxonomic/temporal coverage
- Keywords
- Hosting institution information
- Information regarding associated project(s)
- Sampling methods
- How to cite the dataset
- Museum collection (if applicable)
- Other external links (e.g. a homepage) or additional metadata

We review each of these sections below.

2.4.4.1.1 Title The IPT requires you to provide a *Shortname*. Shortnames serve as an identifier for the resource within the IPT installation (so should be unique within your IPT), and will be used as a parameter in the URL to access the resource via the Internet. Please use only alphanumeric characters, hyphens, or underscores. E.g. *largenet_im* in http://ipt.vliz.be/eurobis/resource?r=largenet_im. After creating a new dataset resource, the field title will be filled out with the short name you provided earlier. Please make sure you provide a dataset title following the guidelines below.

Dataset titles provided to OBIS node managers are often very cryptic, such as an acronym, and often only understandable by the data provider. However, to increase the discoverability and be useful for a larger audience, the dataset title should be as descriptive and complete as possible. OBIS recommends titles to contain information about the taxonomic, geographic and temporal coverage. If the dataset title does not meet these criteria and you believe the title should be changed, then contact the data provider with a suggestion or ask for a more descriptive title. If the dataset has already been published (made publicly available) - and therefore known by that title elsewhere, then the same title should be kept (even if it would not meet the proposed guidelines)! Changing the title of an already published dataset cannot be done, as this will generate confusion and possible duplicates in systems like OBIS or GBIF in a later stage.

The acronym or working title could still be documented in the metadata, so there is no confusion about how the full title is linked to the originally provided acronym or working title.

Caution: Always consult the data provider when changing a dataset title to a more workable and descriptive version.

Originally received title	Title Recommended by Node Manager
BIOCEAN	BIOCEAN database on deep sea benthic fauna
Biomôr	Benthic data from the Southern Irish Sea from 1989-1991
Kyklades	Zoobenthos of the Kyklades (Aegean Sea)
REPHY	Réseau de Surveillance phytoplanctonique

2.4.4.1.2 Abstract The abstract or description of a dataset provides basic information on the content of the dataset. The information in the abstract should improve understanding and interpretation of the data. It is recommended that the description indicates whether the dataset is a subset of a larger dataset and – if so – provide a link to the parent metadata and/or dataset.

If the data provider or OBIS node require bi- or multilingual entries for the description (e.g. due to national obligations) then the following procedure can be followed:

- Indicate English as metadata language
- Enter the English description first
- Type a slash (/)
- Enter the description in the second language

Example: The Louis-Marie herbarium grants a priority to the Arctic-alpine, subarctic and boreal species from the province of Quebec and the northern hemisphere. This dataset is mainly populated with specimens from the province of Quebec. / L'Herbier Louis-Marie accorde une priorité aux espèces arctiques-alpines, subarctiques et boréales du Québec, du Canada et de l'hémisphère nord. Ce jeu présente principalement des spécimens provenant du Québec.

2.4.4.1.3 People and Organizations The EML has several possible roles/functions to describe a contact, creator, metadata provider and associated party.

The `contact` is the person or organization that curates the resource and who should be contacted to get more information or to whom questions with the resource or data should be addressed. Although a number of fields are not required, we strongly recommend providing as much information as possible, and in particular the email address. This will also be the contact information that appears on the OBIS metadata pages.

The `creator` is the person or organization responsible for the original creation of the resource content. When there are multiple creators, the one that bears the greatest responsibility is the resource creator, and other people can be added as associated parties with a role such as 'originator', 'content provider', 'principal investigator', etc.

Possible functions/roles:

- Originator (person/organization that originally gathered/prepared the dataset)
- Content provider (principal person/organization that contributed content to the dataset)

If the resource contact and the resource creator are identical, the IPT allows you to easily copy the information.

The `metadata provider` is the person or organization responsible for producing the resource metadata. If the metadata are provided by the original data provider, then his/her contact details should be filled in. If no metadata are available (e.g. for historical datasets, with no contact person), then the metadata can be completed by e.g. the OBIS node manager and the OBIS node manager becomes the metadata provider.

The `Associated Parties` contains information about one or more people or organizations associated with the resource in addition to those already covered on the IPT Basic Metadata page. For example, if there would be multiple contact persons or metadata creators, they can be added in this IPT section. The principal contact/creator should, however, be added in the IPT Basic Metadata section, not the `Associated Parties` section. It is recommended to complete this section together with the IPT Basic Metadata page, to avoid confusion or overlap in added information.

Possible functions/roles for associated parties are:

- Custodian steward (person/organization responsible for/takes care of the dataset paper)
- Owner (person/organization that owns the data – may or may not be the custodian)
- Point of contact (person/organization to contact for further information on the dataset)
- principal investigator (primary scientific contact associated with the dataset)

Notes:

The owner of a dataset will, in most cases, be an institute, and not an individual person. Although the fields 'last name', and 'position' are indicated as mandatory fields, it is possible to just add the institute name in the 'last name' field for the role 'owner'.

The contact persons in the metadata (contact, creator, metadata creator) are used in the dataset citation (auto-generation) and those added as 'associated parties' are not included as "co-authors".

2.4.4.1.4 License and IP Rights OBIS has published its guidelines on the sharing and use of data here. The recommended licenses for datasets published in OBIS are the Creative Commons Licenses (CC-0, CC-BY, CC-BY-NC), of which CC-0 is the most preferred and CC-BY-NC is least preferred. A Creative Commons license means:

- You are free:
  - to share => to copy, distribute and use the database
  - to create => to produce works from the database
  - to adapt => to modify, transform and build upon the database

In case of CC-0: `public domain`: CC-0 is the preferred option identified by the OBIS steering group. You waive any copyright you might have over the data(set) and dedicate it to the public domain. You cannot be held liable for any (mis)use of the data either. Although CC-0 doesn't legally require users of the data to cite the source, it does not take away the moral responsibility to give attribution, as is common in scientific research. A good blog on why using CC-0 can be found here.

In case of CC-BY: `Attribution`: You must attribute any public use of the database, or works produced from the database, in the manner specified in the license. For any use or redistribution of the database, or works produced from it, you must make clear to others the license of the database and keep intact any notices on the original database.

In case of CC-BY-NC: `non-commercial`: like CC-BY but commercial use is not allowed. This licence can be problematic when the data is re-used in scientific journals.

2.4.4.1.5 Coverage

2.4.4.1.5.1 Geographic Coverage The IPT allows you to enter the geographic coverage by dragging the markers on the given map or by filling in the coordinates of the bounding box. In the description field, a more elaborate text can be provided to describe the spatial coverage indicating the larger geographical area where the samples were collected. For the latter, the sampling locations can be plotted on a map and – by making use of a Gazetteer – the wider geographical area can be derived: e.g. the relevant Exclusive Economic Zone (EEZ), IHO, FAO fishing area, Large Marine Ecosystem (LME), Marine Ecoregions of the World (MEOW), etc. The Marine Regions' Gazetteer might prove to be a useful online tool to define the most relevant sea area(s). There are also LifeWatch Geographical Services that translate geographical positions to these wider geographical areas.

The information given in this section can also help the OBIS node manager in geographic quality control. If the geographic coverage in the EML e.g. is "North Sea", but a number of data points are outside of this scope, then this may indicate errors, and should be checked with the data provider.

If the dataset covers multiple areas (e.g. samples from the North Sea and the Mediterranean Sea), then this should clearly be mentioned in the `geographicDescription` field. Note that the IPT only allows one bounding box, and you have to uncheck the "Set global coverage" box to change box bounds.

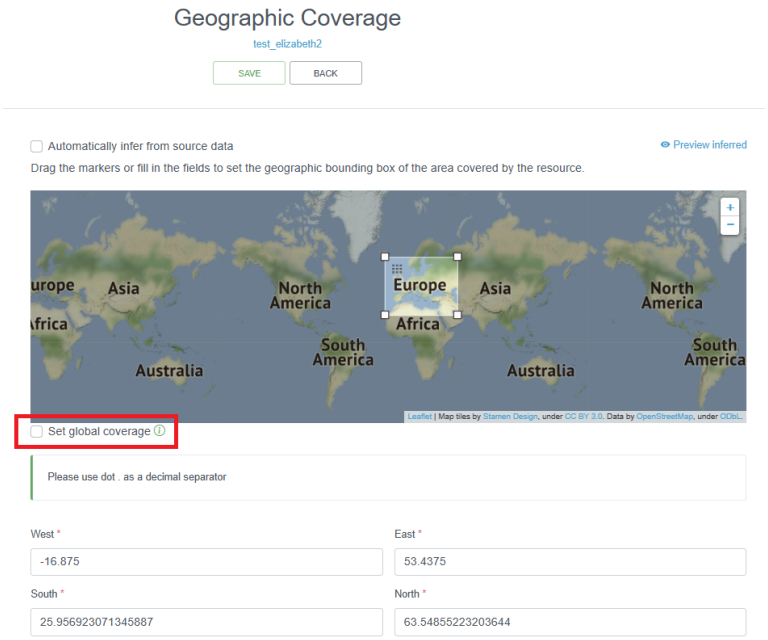


Figure 2.3: Screenshot of the Geographical Coverage section of the metadata, emphasizing how to change the bounds of the coverage box in the map.

2.4.4.1.5.2 Taxonomic Coverage This section can capture two things:

1. A description of the range of taxa that are addressed in the data set. OBIS recommends to only add the higher classification (Kingdom, Class or Order) of the involved groups (e.g. Bivalvia, Cetacea, Aves, Ophiuroidea...). You can easily draw a list of higher taxonomic ranks from the WoRMS taxon match service (or ask the data provider). The taxonomic coverage is not a mandatory field, but the information stored here can be very useful as background information. The description can also contain common names, such as e.g. benthic foraminifera or mussels.
2. An overview of all the involved taxa (not recommended, as all the taxa are already listed in the dataset).

   Note: OBIS also recommends to add information on the (higher) taxonomic groups in the (descriptive) dataset title and abstract.

2.4.4.1.5.3 Temporal Coverage The temporal coverage will be a date range, which can easily be documented. If it is a single date, the start and end date will be the same. The information added here can be used as a quality check for the actual dates in the datasets.

You can also document the Formation Period or the Living Time Period in this section for specimens that may not have been alive during the collection period, or to indicate the time during which the collection occurred.

2.4.4.1.6 Keywords Relevant keywords facilitate the discovery of a dataset. An indication of the represented functional groups can help in a general search (e.g. plankton, benthos, zooplankton, phytoplankton, macrobenthos, meiobenthos ...). Assigned keywords can be related to taxonomy, habitat, geography or relevant keywords extracted from thesauri such as the ASFA thesaurus, the CAB thesaurus or GCMD keywords.

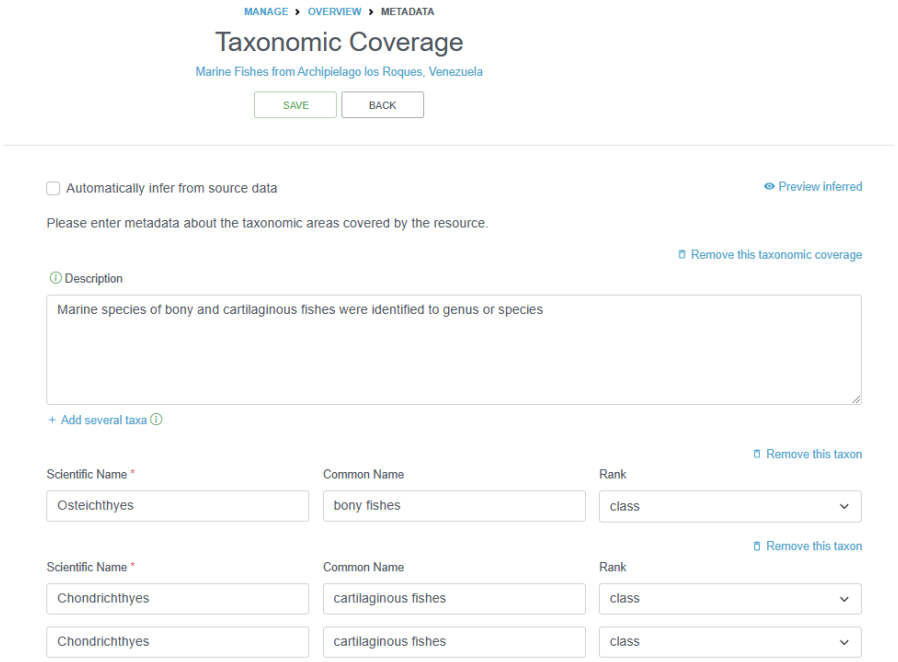


Figure 2.4: Example of the Taxonomic Coverage section of the metadata

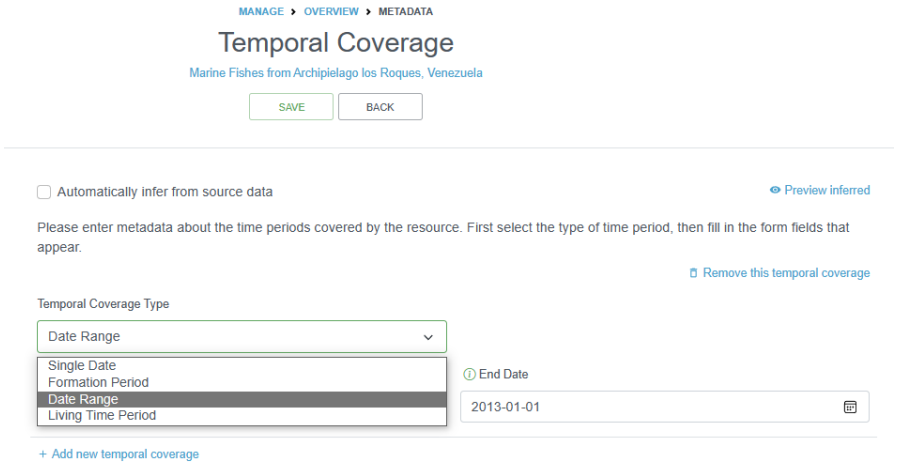


Figure 2.5: Example of the Temporal Coverage section of the metadata

As taxonomy and geography are already covered in previous sections, there is no need to repeat related keywords here. Please consult your data provider which (relevant) keywords can be assigned.

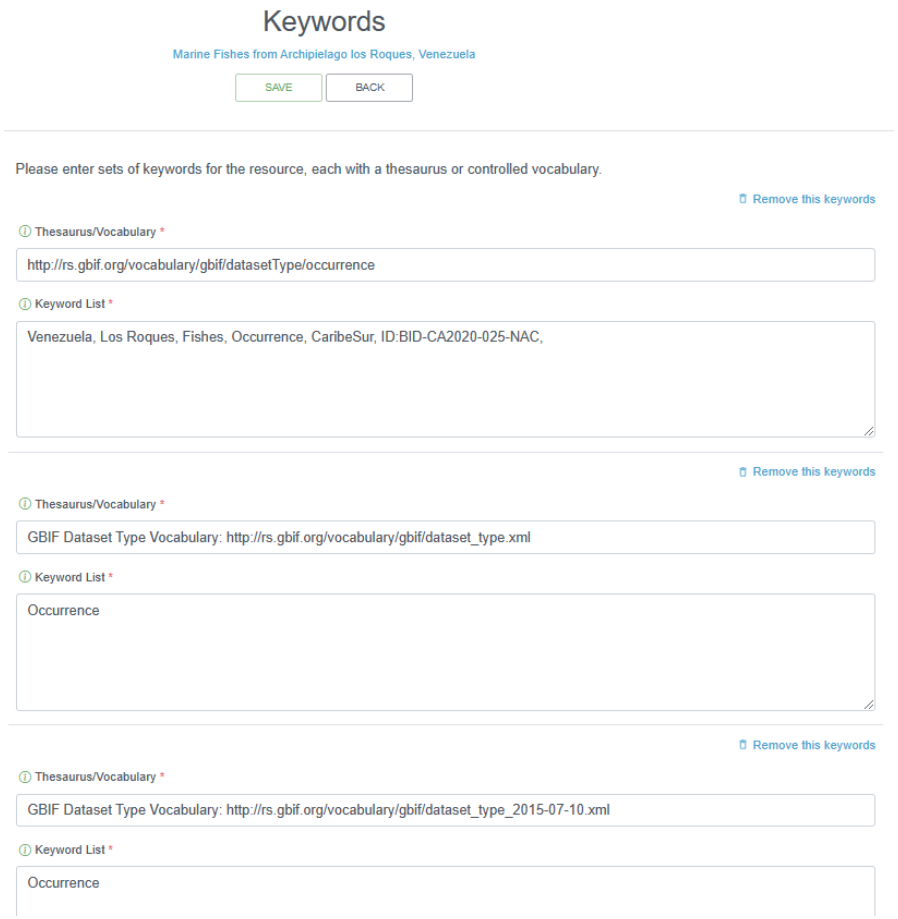


Figure 2.6: Example of the Keywords section of the metadata, showing input for a marine fishes dataset

2.4.4.1.7 Project If the dataset in this resource is produced under a certain project, the metadata on this project can be documented here. Part of the information entered here, can partly overlap with information given in other sections of the metadata (e.g. study area description can have lot of parallel with the geographic coverage section). Personnel involved in the project can be documented or repeated here as well. This is not a problem.

2.4.4.1.8 Sampling Methods The EML can contain descriptions of the sampling and data processing methods. Study extent can be documented here as well to report a more specific geographic area as well as the sampling frequency. Descriptions of sampling procedures, quality control, and steps (sample or data processing) can be given in the same way as the methods section of a scientific paper.

Note that OBIS best practice is to add sampling facts to the extended MeasurementorFact extension, linked to the sampling events in the Event core via eventID.

2.4.4.1.9 Citations The dataset citation allows users to properly cite your dataset in further publications or other uses of the data. When users download datasets from the OBIS download function, a list of the dataset citations packaged with the data in a zipped file is provided.

A dataset citation is different from the data source citation (in case the data is digitized from a publication), and these references can be added to the additional metadata (see bibliography below). A dataset citation can have the same format of a journal article citation, and should include the authors (contact, creator, principle investigator, data managers, custodians, collectors. . .), the title of the dataset, the name of the data publisher (or custodian institute), and the access point URL to the resource.

GBIF's IPT has an auto-generation - Turn On/Off - tool to let the IPT auto-generate the resource citation for you. The citation includes a version number, which is especially important for datasets that are continuously updated. The dataset citation can also include a Citation Identifier - a DOI, URI, or other persistent identifier that resolves to an online dataset web page.

The OBIS node data managers should try to implement a certain degree of format standardization for the dataset citations. The IPT provides an option to auto-generate a citation based on the EML and is formatted as follows: {dataset.authors} ({dataset.pubDate}) {dataset.title}. [Version {dataset.version}]. {organization.title}. {dataset.type} Dataset {dataset.doi}, {dataset.url}

2.4.4.1.10 Bibliography The EML can include the citation of the publications that are related to the described dataset. They can describe the dataset, be based on the dataset or be used in this dataset. Publications can be scientific papers, reports, PhD or master theses. If available, the citation should include the DOI at the end.

This overview will contribute to a better understanding of the data as these publications can hold important additional information on the data and how they were acquired.

2.4.4.1.11 Collection Data This IPT section should only be filled out if there are specimens held in a museum. If relevant, it is strongly recommended that this information is supplied by the data provider or left blank. The collection name, specimen preservation method, and curatorial units should be provided, as applicable.

2.4.4.1.12 External Links This section can include URLs to the resource homepage, to download or find additional information. You can also provide links to your resource if it is hosted elsewhere in different formats.

Links to the online dataset on the OBIS website can be added once the data is available there. For these OBIS links, the required fields should be completed as follows:

- Name: online dataset
- Character set: UTF-8
- Data format: html

If other links are added, then the data format for web-based data is 'html'. If the link refers to a file, the data format of the file will need to be added (e.g. .xlsx, .pdf . . .). The character set for all Darwin Core files is UTF-8, whereas for other web pages this can vary, so you may need to confirm.

2.4.4.1.13 Additional Metadata Any remaining information that could not be catalogued under any of the other metadata, can be mentioned here. This may include logos, purpose of the dataset, a description of how the dataset will be maintained, etc.

2.5 OBIS nodes

Note the OBIS node TOR and system architecture is currently under review and will be updated after the 2023 Steering Group meeting. The information below may change.

OBIS Nodes are either national projects, programmes, institutes, or organizations, National Ocean Data Centers or regional or international projects, programmes and institutions or organizations that carry out data management functions.

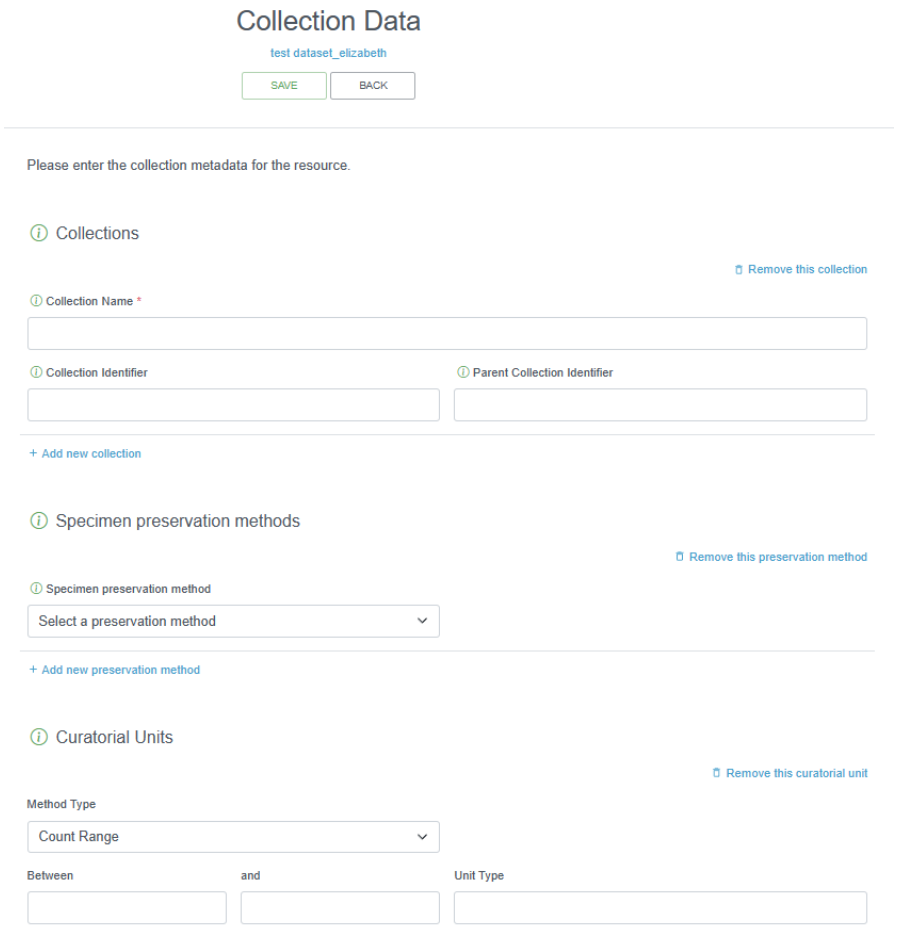


Figure 2.7: Screenshot of the Collection Data page showing what information can be provided for museum specimens

OBIS nodes are responsible for **representing all aspects of OBIS within a particular region or taxonomic domain**. Additional responsibilities include:

- Establishing relationships with key data providers within their geographical (or taxonomic) area of responsibility
- Bringing data and corresponding metadata into the global database to be shared with the OBIS community
- Responsibility for **all aspects of the data**
- Gaining permission to providing access to the data
- Ensuring a certain level of data quality
- Transfer of these datasets to the global OBIS database
- Provide support for the full implementation of OBIS worldwide by serving on the IODE Steering Group for OBIS and any relevant Task Teams or ad hoc project teams
- Each node may also maintain a data presence on the Internet representing their specific area of responsibility

2.5.1 Terms of Reference of OBIS nodes

Data Responsibilities

- Receiving or harvesting marine biodiversity data (and metadata) from national, regional, and international programs, and the scientific community at large, and from Tier III nodes by Tier II nodes, and from Tier II nodes by Tier I nodes
- Perform data validation (using standards, tools, and best practices), as described in the OBIS manual (Tier II)
- Reporting the results of quality control directly to data collectors/originator (or Tier III node) as part of the quality assurance activity
- Making data (and metadata) available to OBIS using agreed upon standards and formats which are described in the OBIS Manual (Tier II), making data available to Tier II nodes (Tier III)
- Control data access, terms of use and sharing policies
- Comply with the IOC/OBIS data policy for using and sharing OBIS data
- Contribute to the development of standards and best practices in OBIS (recommended)
- Contribute to the development of open-source tools in OBIS (recommended)
- Ensuring the long-term preservation of the data, metadata and associated information required for correct interpretation of the data (including version-control) (recommended)
- Build customized data portals (optional)

Administration Responsibilities

- Become a member of the IODE steering group for OBIS, attend the SG-OBIS annual meeting and report on node activities
- Provide indicators on up-time, responsiveness, and data processed by nodes and present a report to SG-OBIS
- Customer support (data queries, analyses, feedback)
- Outreach and Capacity Building (i.e., providing expertise, training and support in data management, technologies, standards and best practices)
- Engage in stakeholder groups (recommended)

2.5.2 How to become an OBIS node

OBIS nodes now operate under the IODE network as either National Oceanographic Data Centres (NODCs) or Associate Data Unites (ADUs). Prospective nodes are required to apply to the IODE for membership.

The procedure to become an OBIS node is as follows:

- If you are an existing NODC (within the IODE network) and the OBIS node activities fall under the activities of the NODC:

  - Send a letter expressing your interest to become an OBIS node (including contact information of the OBIS node manager, and geographical/thematic scope of your OBIS node)
- If you are not an existing NODC:
  - Email your application form to become an IODE Associate Data Unit (ADU), with a specific role as OBIS node. Applications for ADU membership in OBIS shall be reviewed by the IODE Officers in consultation with the IODE Steering Group for OBIS.

2.5.3 OBIS Node Health Status Check and Transition Strategy

OBIS nodes should operate under IODE as either IODE/ADU or IODE/NODC. As such OBIS nodes are a member of the IODE network.

The IODE Steering Group (SG) for OBIS evaluates the health status of OBIS nodes at each annual SG meeting, and considers an OBIS node as **inactive** when it meets any of the following conditions:

1. The OBIS node manager recurrently fails to answer the communications from the project manager or the SG co-chairs in the last 12 months
2. The OBIS node manager or a representative fails to attend (personally or virtually) the last 2 SG meetings without any written reason
3. The OBIS node does not have an IPT
4. The OBIS node has an IPT, but it has not been running for the last 12 months
5. The datasets in the OBIS node's IPT have been removed and not restored in the last 12 months (without any explanation)
6. The OBIS node has not provided new data for the last 2 years

The OBIS Secretariat prepares a health status check report of each OBIS node based on the six items above and informs the OBIS node manager on their status 3 months before the SG meeting. At the SG meeting, the SG-OBIS co-chair will present the results of the OBIS nodes health status check report including a listing of the inactive OBIS nodes. The SG-OBIS members representing active OBIS Nodes will make one of the following decisions:

1. Request the inactive OBIS node to submit a plan with actions, deliverables and times to improve their performance, within 3 months, to the OBIS Secretariat. This plan is reviewed and accepted by the OBIS-Executive Committee Or
2. Provide a recommendation to the IOC Committee on IODE to remove the OBIS node from the IODE network.

In either case, the OBIS Secretariat will inform the OBIS node manager of the SG-OBIS decision, with a copy to the IODE officers and the IODE national coordinator for data management of the country concerned.

The IODE Committee is requested to consider the recommendation from the OBIS Steering Group and it may either accept the recommendation or request the inactive OBIS node to submit an action plan (option 1).

When the inactive OBIS node is removed from the IODE network, the SG-OBIS will ask whether another OBIS node is interested in taking over the responsibilities of the removed OBIS node, until a new OBIS node in the country/region is established.

Data Formatting

Chapter 3

Dataset structure

Formatting data can be challenging. This section of the manual deals with how to format data for OBIS, beginning with an overview of dataset structure.

Deciding on your dataset structure is one of the first steps towards getting your data ready for publishing. At this step, there are different non arbitrary you need to do with your data, but it is important to determine which structure best suits your dataset before proceeding. Then, once you have decided on the dataset structure, you can continue formatting your data.

We have created the following flow chart for an overview on how to determine what structure best suits your data.

What kind of data do you have, or will collect?

What kind of data do...

occurrence

occurrence

genetic

genetic

biomass

biomass

tracking

tracking

habitat

habitat

Do you have event level information (e.g. sample, station, cruise, study, etc.)?

Do you have event level infor...

abundance, percent cover

abundance, perce...

Yes

Yes

No

No

Have (or will) you collect any data associated with samples or sampling? (e.g. temperature, length, etc.)

Have (or will) you collec...

Event core

Event core

Occurrence Core

Occurrence Core

Yes

Yes

No

No

Event core

Event core

Event core only

Event core...

Occurrence Core

Occurrence Core

Have (or will) you collect any data associated with samples or sampling? (e.g. temperature, length, etc.)

Have (or will) you collec...

Yes

Yes

No

No

Occurrence core

Occurrence...

Occurrence core

Occurrence...

DNAderived dataextension

DNA...

Start

Start

eMoFextension

eMoF...

eMoFextension

eMoF...

Have (or will) you collect any data associated with samples or sampling? (e.g. temperature, length, etc.)

Have (or will) you collec...

Yes

Yes

No

No

Occurrencecore

Occurrence...

Occurrencecore only

Occurrence...

eMoFextension

eMoF...

DNAderived dataextension

DNA...

DNAderived dataextension

DNA...

Text is not SVG - cannot display

For more guidance, see the sections below.

3.1 When to use Event Core

Event Core describes **when** and **where** a specific sampling event happened and contains information such as location and date. Event Core is often used to organize your data tables when there are more than one sampling occasion and/or location, and different occurrences linked to each sampling. This organization follows the rationale of most ecological studies and typical marine sampling design. It covers:

- When specific details are known about **how** a biological sample was taken and processed. These details can then be defined in the eMoF Extension with the Q01 vocabulary
- When the dataset contains abiotic measurements, or other biological measurements which are **related to an entire sample** (not a single specimen). For example a biomass measurement for an entire sample, not each species within the sample

Event Core can be used in combination with the Occurrence and eMoF extensions. The identifier that links Event Core to the extension is the `eventID`. `parentID` can also be used to give information on hierarchical sampling. `occurrenceID` can also be used in datasets with Event Core in order to link information between the Occurrence extension and the eMoF extension.

3.2 When to use Occurrence Core

Occurrence Core datasets describe **observations** and **specimen records** and cover instances when:

- **No information** on how the data was sampled or how samples were processed is available
- No abiotic measurements are taken or provided

- You have eDNA and DNA-derived data
- Biological measurements are made on **individual specimens** (each specimen is a single occurrence record)

Occurrence Core is also often the preferred structure for museum collections, citations of occurrences from literature, and sampling activities.

Datasets formatted in Occurrence Core can use the eMoF Extension for when you have biotic measurements or facts about your specimen. The DNA derived data extension can also be used to link to DNA sequences. The identifier that links Occurrence Core to the extension(s) is the `occurrenceID`.

3.3 Extensions in OBIS

Currently OBIS accepts the following extensions:

- Occurrence
- Event
- MeasurementOrFact
- extendedMeasurementOrFact
- DNADerivedData

3.3.1 How are extensions linked to core tables in OBIS?

As established in the relational database section, OBIS relies on datasets being formatted according to a relational database structure. The ENV-DATA approach that OBIS implements means your dataset will have a Core table and (optionally) Extension tables. As a review, a core file contains information relevant and applicable to each record in the extension(s). An extension file then contains records that link back to a record in the core file with more specific information (e.g., methods, measurements, facts, DNA sequences, etc.).

The extension file(s) accepted by OBIS (eMoF, Occurrence, DNA) are linked to your core tables by the use of identifying ID codes. These codes could be either `eventID` or `occurrenceID`. For details on how to construct these IDs, click here.

3.3.2 Differences between identifiers

If your core file is based on occurrences (e.g., a record of one or more taxa specimens), then any extensions are linked with `occurrenceID`. If your core file is based on events (e.g., a sampling event, cruise, observation, etc.), then the linking identifier is `eventID`. In the Core tables, identifiers are always unique, which means, they do not repeat and each line has a different identifier. On the other hand, multiple records in an extension file can have the same identifier which will link them to the same event or occurrence record (depending on which is the Core). The different linking identifiers are shown in the figure below.

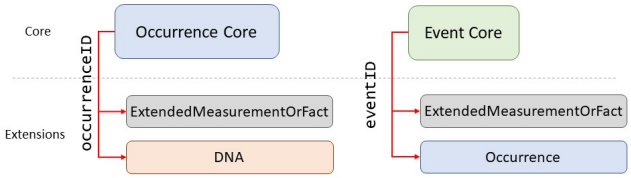


Figure 3.1: Diagram of how the different core tables are linked to their extensions by different identifiers.

Let us consider a fictional plankton trawl sampling event to demonstrate how identifiers link Core and Extension tables in OBIS. This trawl used two types of nets, occurred in March 2013, and has an eventID

`plankton-northsea-2013-03`. Suppose we have information about the types of trawl used and the species abundance from this trawling event. The information (e.g., date) of the sampling event itself would be found in the Event Core, whereas the abundance data and sampling methods would be in the eMoF table. How do we ensure the abundance and sampling method data is properly linked to the correct event? By using the same eventID for each record in the eMoF table, `plankton-northsea-2013-03`, the information is properly linked between the Event Core and the eMoF extension.

3.4 Data formatting tools

The GBIF Norwegian Node created the DwC Excel Template Generator. This tool will generate four different types of blank Excel spreadsheets: Occurrence Core, MeasurementOrFact, Metadata, and a README. This tool works best if you already know which Darwin Core fields you need, 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 that helps create templates for Event (aka Sampling-Event) and Occurrence core tables, as well as MeasurementsOrFacts, Extended MeasurementsOrFacts, and Simple Multimedia extensions. GBIF provides an Occurrence core template and an Event core template. If you use these templates from GBIF, be aware that GBIF's required terms are different from OBIS.

There are also some tools that can help you unpivot (or flatten) data tables. These can be used to flatten many columns into one, particularly useful for the eMoF table.

- GBIF Norway's crosstab to list converter. Note that this tool is not completely automated
- Excel's built-in unpivot function

3.5 Constructing and using indentifier codes

Content

3.5.1 eventID

Using a unique identifier for each physical sample or subsample in your dataset taken at each location and time is highly recommended to ensure sample traceability and data provenance. `eventID` is an identifier for an individual sampling or observation event, whereas `parentEventID` is an identifier for a parent event, which is composed of one or more sub-sampling (child) events (`eventIDs`).

`eventID` can be used for replicated samples or sub-samples. It is important to make sure each replicate sample receives a unique `eventID`, which could be based on the unique sample ID in your dataset. Sample ID can also be recorded in `materialSampleID`, as OBIS does not need to have separate `eventID`s and `materialSampleIDs`. Rather OBIS can treat these two terms as equivalent. Be sure to still fill in the `eventID` field if you want to use `materialSampleID`, as OBIS only uses `eventID` and `parentEventID` for structuring datasets, not sample ID. This does not prevent you from using the field if you would like to.

If you do not already have a `materialSampleID`, creating a unique `eventID` for your data records can be as straightforward as combining different fields from your data.

> **Note** You should consider carefully what combination of fields will generate a **unique** event. Combinations including date, time, location, and depth are common elements to help generate such unique codes.

Including the event type can also be useful for datasets with hierarchical sampling methods (e.g., samples taken from a station within a cruise). Repeating the `parentEventID` in the child event (use `:` as delimiter) can make

the structure of the dataset easier to understand. Nesting event information in this way also allows you to reduce redundancy and still provide information relevant to each level of sampling.

Broadly, an `eventID` can take the form of `[parentEventID]:[sample type]_[sample ID]`

Thus to construct a unique `eventID` for parent and child events, you join relevant sampling information. Possible configurations (with examples) could include:

- Project_cruise_station_date_sample
  - STAR_arcticsea_st3520_1989-04-04_s01
- Project_habitat_Genus_species_year_sampletype_samplenumber
  - BEE_seamount_Genus_species_2013_cruise_s123
- Institution_year_location_samplemethod_sample
  - Concordia_2003_Coast_Station1_seine_s01
  - Concordia_2003_Coast_Station1_trap_s01

These examples are not exhaustive and other similarly structured variations that fit your data are acceptable. Consider also including year within your `eventID`s to ensure codes remain globally unique in subsequent years, which is particularly useful if your sampling protocol is repeated temporally. Remember, what is the main information about a sampling event that helps you identify it? For instance, it is helpful when we know the location, date, project, habitat. So you can build your `eventID` code based on this information and ensuring they will not repeat (e.g., will result in a unique identifier).

Information related to your sampling events can be assigned to the highest relevant event level in order to avoid repetition of information. For example, if all samples taken from a station occurred at the same depth, this information can be listed once. Variation between samples (e.g., exact time or coordinates) can also be easily reflected for each event. See the table below for a demonstration.

eventID	parentEventID	eventRemarks	eventDate	maximumDepthInMeters
cruise_1		cruise		
cruise_1:station_1	cruise_1	station		15
cruise_1:station_1:core_1	cruise_1:station_1	sample	2011-03-06T08:35	
cruise_1:station_1:core_2	cruise_1:station_1	sample	2011-03-06T08:52	
cruise_1:station_1:core_1:subsample_1	cruise_1:station_1:core_1	subsample		

We recommend using controlled vocabulary for the "type" column. Although no standards have been agreed upon yet, commonly used terms for event `type` included are `cruise`, `stationVisit`, `transect`, `quadrat`, `sample`, `subSample`.

Consider another example from a real dataset below:

eventID	parentEventID	eventDate	eventRemarks
IOF_benthos_Plominski_zaljev_2000_crs			cruise
IOF_benthos_Plominski_zaljev_2000_stat1	IOF_benthos_Plominski_zaljev_2000_crs	2000-08	stationVisit
IOF_benthos_Plominski_zaljev_2000_stat2	IOF_benthos_Plominski_zaljev_2000_crs	2000-08	stationVisit
IOF_benthos_Plominski_zaljev_2000_s01	IOF_benthos_Plominski_zaljev_2000_stat1		sample
IOF_benthos_Plominski_zaljev_2000_s02	IOF_benthos_Plominski_zaljev_2000_stat2		sample

Data from Environmental impact assessments in the eastern part of Adriatic sea - species list of benthic invertebrates and phytobenthos (2000-2010).

We can see that each record has a similar `eventID` structure, except for the last part which indicates the event type - documented in the `eventRemarks` column. In this dataset, records with the `eventID` `IOF_benthos_Plominski_zaljev_2000_crs` has information applicable for records with `eventID`s ending with `_stat1`, `_stat2`, `_s01`, and `_s02` because `_crs` is their parent event. Similarly, information (e.g., date of station visit, coordinates) documented in records with `eventID` `IOF_benthos_Plominski_zaljev_2000_stat1` is applicable for the two sample records (`eventID` `_s01` and `_s02`), because these samples were taken at Station 1 (indicated by the `parentEventID`). These `eventID`s could have been nested in another way, such as `IOF_benthos_Plominsku_zaljev_2000_crs:stat1:s01` which would embed the `parentEventID` into the identifier.

See also De Pooter et al. 2017 for an example of an event hierarchy in a complex benthos dataset.

Watch this video for a demonstration on how to construct eventIDs:

3.5.2 occurrenceID

`occurrenceID` is an identifier for occurrence records. Each occurrence record should have a globally unique identifier. Because `occurrenceID` is a required term, you may have to construct a persistent and globally unique identifier for each of your data records if none already exists (e.g., if records were not labeled with unique identifiers before, such as during sample processing or image/sensor detection).

There are no standardized guidelines yet on designing the persistence of this ID, the level of uniqueness (from within a dataset to globally in OBIS), and the precise algorithm and format for generating the ID. But in the absence of a persistent globally unique identifier, one can be constructed by combining the `institutionCode`, the `collectionCode` and the `catalogNumber` (or autonumber in the absence of a catalogNumber). This is similar to how `eventID` is constructed. You may also follow Life Science Identifiers guidelines. Note that the inclusion of `occurrenceID` is also necessary for datasets in the OBIS-ENV-DATA format.

An important consideration for museum specimens: there is the possibility that the institution a specimen is housed at may change. Therefore you may consider omitting institution identifiers within an `occurrenceID`, because `occurrenceID` should **not** change over time.

See the example below:

modified	institutionCode	collectionCode
2017-02-27 15:47:31	Ugent	Vegetation__Gazi__Bay(Kenya)1987
2017-02-27 15:47:31	Ugent	Vegetation__Gazi__Bay(Kenya)1987
2017-02-27 15:47:31	Ugent	Vegetation__Gazi__Bay(Kenya)1987

basisOfRecord	occurrenceID	catalogNumber
HumanObservation	Ugent__Vegetation__Gazi__Bay(Kenya)1987__7553	Ugent__Vegetation__Gazi__Bay(Kenya)1987__7553
HumanObservation	Ugent__Vegetation__Gazi__Bay(Kenya)1987__7554	Ugent__Vegetation__Gazi__Bay(Kenya)1987__7554
HumanObservation	Ugent__Vegetation__Gazi__Bay(Kenya)1987__7555	Ugent__Vegetation__Gazi__Bay(Kenya)1987__7555

Data from Algal community on the pneumatophores of mangrove trees of Gazi Bay in July and August 1987.

Chapter 4

Formatting data tables

4.1 Darwin Core Term Checklist for OBIS

There are many Darwin Core terms listed in the TDWG quick reference guide. However, not all these terms are necessary for publishing data to OBIS.

For your convenience, we have created a checklist of all the Darwin Core terms relevant for OBIS data providers. You can reference this list to quickly see which terms are required by OBIS, which file (Event, Occurrence, eMoF, DNA) they can be found in, and which Darwin Core class it relates to. These terms correlate with the IPT vocabulary mapping you will do when it comes time to publish your dataset. You may notice some terms are accepted in multiple data tables (e.g., Event and Occurrence) - this is because it depends on your dataset structure. If you have an Event Core, you will include some terms that would not be included if you had Occurrence Core. For guidance on specific class terms (e.g., location, taxonomy, etc.), see the Darwin Core section of the manual.

Note that when you publish your dataset on the IPT, if you use a term not listed below it will be an unmapped field and will **not** be published alongside your data. You may still wish to include such fields in your dataset if you are publishing to other repositories, just know that they will not be included in your OBIS dataset. You may include this information either by putting it in the `dynamicProperties` field in JSON format, or putting the information into the eMoF. Alternatively, you may have fields that you do not wish to be published and that do not correspond to one of these terms (e.g. personal notes). This is okay - if they are not mapped to one of the terms, that column in your dataset will not be published.

Term	OBIS Required	DarwinCore Class	Event	Occurrence	eMoF	DNA
eventDate	required	event	x	x		
eventID	required	event	x	x	x	
decimalLatitude	required	location	x	x		
decimalLongitude	required	location	x	x		
occurrenceID	required	occurrence		x	x	x
occurrenceStatus	required	occurrence		x		
basisOfRecord	required	record		x		x
scientificName	required	taxon		x		
scientificNameID	strongly recommended	taxon		x		
DNA_sequence	strongly recommended	dna				x
env_broad_scale	strongly recommended	dna				x
env_local scale	recommended	dna				x
env_medium	strongly recommended	dna				x
lib_layout	recommended	dna				x
nucl_acid_amp	recommended	dna				x
nucl_acid_ext	recommended	dna				x
otu_class_appr	recommended	dna				x
otu_db	recommended	dna				x
otu_seq_comp_appr	recommended	dna				x
pcr_primer_forward	strongly recommended	dna				x
pcr_primer_name_forward	strongly recommended	dna				x
pcr_primer_name_reverse	strongly recommended	dna				x
pcr_primer_reference	strongly recommended	dna				x
pcr_primer_reverse	strongly recommended	dna				x
samp_name	recommended	dna				x
samp_vol_we_dna_ext	recommended	dna				x
seq_meth	recommended	dna				x
sop	recommended	dna				x

Term	OBIS Required	DarwinCore Class	Event	Occurrence	eMoF	DNA
target_gene	strongly recommended	dna				x
target_subfragment	strongly recommended	dna				x
day	recommended	event	x	x		
endDayOfYear	recommended	event	x	x		
eventRemarks	optional	event	x	x		
eventTime	recommended	event	x	x		
fieldNotes	optional	event	x			
fieldNumber	optional	event	x			
habitat	recommended	event	x		x	
month	strongly recommended	event	x	x		
parentEventID	required (if exists)	event	x			
sampleSizeUnit	strongly recommended	event		x	x	
sampleSizeValue	strongly recommended	event		x	x	
samplingEffort	strongly recommended	event		x	x	
samplingProtocol	strongly recommended	event		x	x	
startDayOfYear	recommended	event	x			
verbatimEventDate	recommended	event	x			
year	strongly recommended	event	x	x		
bed	optional	geologicalContext	x	x		
earliestAgeOrLowestStage	optional	geologicalContext	x	x		
earliestEonOrLowestEonothem	optional	geologicalContext	x	x		
earliestEpochOrLowestSeries	optional	geologicalContext	x	x		
earliestEraOrLowestErathem	optional	geologicalContext	x	x		
earliestPeriodOrLowestSystem	optional	geologicalContext	x	x		
formation	optional	geologicalContext	x	x		
group	optional	geologicalContext	x	x		
highestBiostratigraphicZone	optional	geologicalContext	x	x		
latestAgeOrHighestStage	optional	geologicalContext	x	x		
latestEonOrHighestEonothem	optional	geologicalContext	x	x		
latestEpochOrHighestSeries	optional	geologicalContext	x	x		
latestEraOrHighestErathem	optional	geologicalContext	x	x		
latestPeriodOrHighestSystem	optional	geologicalContext	x	x		
lithostratigraphicTerms	optional	geologicalContext	x	x		
lowestBiostratigraphicZone	optional	geologicalContext	x	x		
member	optional	geologicalContext	x	x		
dateIdentified	optional	identification		x		
identificationID	optional	identification		x		
identificationQualifier	recommended	identification		x		
identificationReferences	optional (required for imaging data)	identification		x		
identificationRemarks	recommended	identification		x		
identificationVerificationStatus	optional (required for imaging data)	identification		x		
identifiedBy	optional (required for imaging data)	identification		x		
identifiedByID	optional	identification		x		
typeStatus	optional	identification		x		
continent	strongly recommended	location	x	x		
coordinatePrecision	strongly recommended	location	x	x		
coordinateUncertaintyInMeters	strongly recommended	location	x	x		
country	recommended	location	x	x		
countryCode	optional	location	x	x		
county	optional	location	x	x		
footprintSpatialFit	optional	location	x	x		
footprintSRS	optional	location	x	x		
footprintWKT	recommended	location	x	x		
geodeticDatum	recommended	location	x	x		
georeferencedBy	optional	location	x	x		
georeferencedDate	optional	location	x	x		
georeferenceProtocol	optional	location	x	x		
georeferenceSources	optional	location	x	x		
higherGeography	optional	location	x	x		
higherGeographyID	optional	location	x	x		
island	optional	location	x	x		
islandGroup	optional	location	x	x		
locality	recommended	location	x	x		
locationAccordingTo	recommended	location	x	x		
locationID	strongly recommended	location	x	x		
locationRemarks	recommended	location	x	x		
maximumDepthInMeters	strongly recommended	location	x	x		
maximumDistanceAboveSurfaceInMeters	optional	location	x	x		
maximumElevationInMeters	optional	location	x	x		
minimumDepthInMeters	strongly recommended	location	x	x		
minimumDistanceAboveSurfaceInMeters	optional	location	x	x		
minimumElevationInMeters	optional	location	x	x		
municipality	optional	location	x	x		
pointRadiusSpatialFit	optional	location	x	x		
stateProvince	optional	location	x	x		
verbatimCoordinates	optional	location	x	x		
verbatimCoordinateSystem	optional	location	x	x		
verbatimDepth	optional	location	x	x		
verbatimElevation	optional	location	x	x		
verbatimLatitude	optional	location	x	x		
verbatimLocality	optional	location	x	x		
verbatimLongitude	optional	location	x	x		
verbatimSRS	optional	location	x	x		
waterBody	recommended	location	x	x		
materialSampleID	recommended	materialSample		x		
measurementAccuracy	recommended	measurementOrFact			x	
measurementDeterminedBy	optional	measurementOrFact			x	
measurementDeterminedDate	optional	measurementOrFact			x	
measurementID	recommended	measurementOrFact			x	
measurementMethod	recommended	measurementOrFact			x	
measurementRemarks	recommended	measurementOrFact			x	
measurementType	strongly recommended	measurementOrFact			x	
measurementTypeID	strongly recommended	measurementOrFact			x	
measurementUnit	strongly recommended	measurementOrFact			x	
measurementUnitID	strongly recommended	measurementOrFact			x	
measurementValue	strongly recommended	measurementOrFact			x	
measurementValueID	strongly recommended	measurementOrFact			x	
associatedMedia	recommended	occurrence		x		
associatedReferences	optional	occurrence		x		
associatedSequences	recommended	occurrence		x		

Term	OBIS Required	DarwinCore Class	Event	Occurrence	eMoF	DNA
associatedTaxa	optional	occurrence		x		
behavior	recommended	occurrence		x	x	
catalogNumber	recommended	occurrence		x		
disposition	optional	occurrence		x		
establishmentMeans	optional	occurrence		x		
georeferenceVerificationStatus	recommended	occurrence		x		
individualCount	strongly recommended	occurrence		x	x	
lifeStage	recommended	occurrence		x	x	
occurrenceRemarks	recommended	occurrence		x		
organismQuantity	strongly recommended	occurrence		x	x	
organismQuantityType	strongly recommended	occurrence		x	x	
otherCatalogNumbers	optional	occurrence		x		
preparations	optional	occurrence		x		
recordedBy	recommended	occurrence		x		
recordedByID	recommended	occurrence		x		
recordNumber	recommended	occurrence		x		
reproductiveCondition	recommended	occurrence		x		
sex	recommended	occurrence		x	x	
associatedOccurrences	optional	organsim		x		
associatedOrganisms	optional	organsim		x		
organismID	recommended	organsim		x		
organismName	recommended	organsim		x		
organismRemarks	recommended	organsim		x		
organismScope	optional	organsim		x		
previousIdentifications	recommended	organsim		x		
accessRights	recommended	record	x	x		
bibliographicCitation	recommended	record	x	x		
collectionCode	optional	record	x	x		
collectionID	optional	record	x	x		
dataGeneralizations	optional	record	x	x		
datasetID	recommended	record	x	x		
datasetName	recommended	record	x	x		
dynamicProperties	recommended	record	x	x		
informationWithheld	optional	record	x	x		
institutionCode	optional	record	x	x		
institutionID	optional	record	x	x		
language	recommended	record	x	x		
license	recommended	record	x	x		
modified	recommended	record	x	x		
ownerInstitutionCode	optional	record	x	x		
references	recommended	record	x	x		
rightsHolder	recommended	record	x	x		
type	strongly recommended	record	x	x	x	
acceptedNameUsage	recommended	taxon		x		
acceptedNameUsageID	recommended	taxon		x		
higherClassification	recommended	taxon		x		
infraspecificEpithet	recommended	taxon		x		
nameAccordingToID	recommended	taxon		x		
namePublishedInID	optional	taxon		x		
namePublishedInYear	optional	taxon		x		
nomenclaturalCode	optional	taxon		x		
nomenclaturalStatus	optional	taxon		x		
originalNameUsage	recommended	taxon		x		
originalNameUsageID	recommended	taxon		x		
parentNameUsage	recommended	taxon		x		
parentNameUsageID	recommended	taxon		x		
phylum	recommended	taxon		x		
scientificNameAuthorship	recommended	taxon		x		
specificEpithet	recommended	taxon		x		
subgenus	recommended	taxon		x		
taxonConceptID	optional	taxon		x		
taxonID	optional	taxon		x		
taxonomicStatus	optional	taxon		x		
taxonRank	strongly recommended	taxon		x		
taxonRemarks	recommended	taxon		x		
verbatimTaxonRank	recommended	taxon		x		
vernacularName	recommended	taxon		x		
type or eventType	strongly recommended	event	x			
class	recommended	taxon		x		
family	recommended	taxon		x		
genus	strongly recommended	taxon		x		
kingdom	strongly recommended	taxon		x		
order	strongly recommended	taxon		x		

4.2 Name Matching Strategy for taxonomic quality control

OBIS requires all your specimens to be classified and matched against an authoritative taxonomic register. This effectively attaches unique stable identifiers (and digitally traceable) to each of your species. Meaning, if a taxonomic ranking or a species name changes in the future, there will be no question as to which species your dataset is actually referring to. Matching to registers also helps to avoid misspelled or unused terms.

OBIS currently accepts identifiers from **three** authoritative lists:

- World Register Marine Species (WoRMS) LSIDs
- Integrated Taxonomic Information System (ITIS) TSNs
- Barcode of Life Data Systems (BOLD) and NCBI identifiers

The identifiers (LSID, TSN, ID) from these registers will be used to populate the `scientificNameID` field. OBIS can accept other LSIDS besides WoRMS, as long as they are mapped in WoRMS. If you would

like to include multiple identifiers, please use a concatenated list where each register is clearly identified (e.g. urn:lsid:itis.gov:itis_tsn:12345, NCBI:12345, BOLD:12345).

> **Note** You should prioritize using LSIDs because they are unique identifiers that indicate the authority the ID comes from.WoRMS LSIDs are also the taxonomic backbone that OBIS relies on, as it is built on marine systems and is linked to the other taxonomic authoritative lists.

You can also use the Interim Register of Marine and Nonmarine Genera (IRMNG) to distinguish marine genera from freshwater genera.

4.2.1 Taxon Matching Workflow

The OBIS node managers have agreed to match all the scientific names in their datasets according to the following Name Matching workflow:

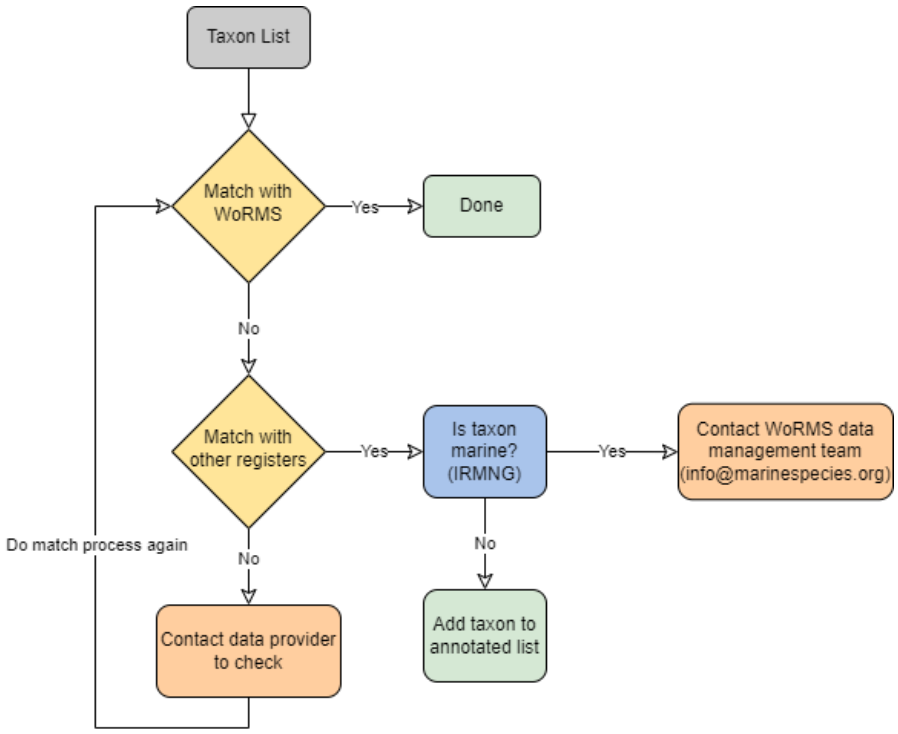


Figure 4.1: Workflow for matching a list of taxon names to WoRMS

4.2.1.1 Step 1: Match with WoRMS

The procedure for matching to WoRMS and then attaching successful matches back to your data can be simplified to:

- Prepare a file (.csv, .txt, .xlsx, etc.) with the list of your specimens/taxa
- Upload the file to WoRMS taxon match tool
  - Check relevant boxes
- Review returned file
- Resolve any ambiguous matches
- Download file and identify data to include in your Occurrence data table for OBIS
  - LSIDs, taxonomic fields, etc.
- Attach LSIDs back to your data using e.g.:

- R (merge)
- Excel (vlookup)

The taxon match tool of the World Register of Marine Species (WoRMS) is an automatic way to download the taxonomic information about your occurrence records, without having to look for each name in the site. It is available at http://www.marinespecies.org/aphia.php?p=match. The WoRMS taxon match will compare your taxon list to the taxa available in WoRMS. The following video demonstrates the basic steps for using the WoRMS Taxon match.

The taxon match takes into account exact matches and fuzzy matches. Fuzzy matches include possible spelling variations of a name available in WoRMS. WoRMS also identifies ambiguous matches, indicating that several potential matching options are available (e.g. homonyms). You can check these ambiguous matches and select the correct one, based on e.g., the general group information (a sponge dataset) or the authority. If this would be impossible with the available information (e.g., missing authority or very diverse dataset), then you need to contact the data provider for clarification. Watch the video below for a demonstration on how to resolve ambiguous or fuzzy matches.

For performance reasons, the limit is set to 1,500 rows for the taxon match tool. Larger files can be sent to info@marinespecies.org and will be returned as quickly as possible. In case you have recorded a taxon that is not registered in WoRMS (e.g., newly discovered species), you should contact them so the database can be updated.

After matching, the tool will return you a file with the AphiaIDs, LSIDs, valid names, authorities, classification, and any other output you have selected.

Note The WoRMS LSID is used for `DwC:scientificNameID`.

A complete online manual is available at http://www.marinespecies.org/tutorial/taxonmatch.php. You can attach IDs obtained from WoRMS back to your own data using Excel's vlookup function. R script to do this is shown below.

R script for attaching Taxon Lists to ID Lists:

If you are familiar enough with R, you can use the `merge` function to attach the two lists to your data. We provide a short example of how to use this function below.

```
#Generate example data table with species occurences, for this example we will only have one column with
data<-data.frame(scientificName=c("Thunnus thynnus", "Rhincodon typus", "Luidia maculata","Ginglymostoma

# this would be your matched file from WoRMS, but for example we are generating a simple list with the s
lsids<- data.frame(scientificName=c("Ginglymostoma cirratum","Luidia maculata","Thunnus thynnus", "Rhinc
 LSID = c("urn:lsid:marinespecies.org:taxname:105846", "urn:lsid:marinespecies.org:taxname:213112","urn:

#merge data frames together
matched_data<-merge(data, lsids, by = "scientificName")
matched_data
```

4.2.1.2 How to fetch a full classification for a list of species from WoRMS?

When setting up your WoRMS taxon match, to obtain the full classification for your list of species, simply check the box labeled "Classification". This will add classification output in addition to the requested identifiers to your taxon match file, including Kingdom, Phylum, Class, Order, Family, Genus, Subgenus, Species, and Subspecies.

4.2.1.3 What to do with non-matching names?

If your scientificName does not find an exact match to the WoRMS database, you may get an **ambiguous** match. According to WoRMS guidelines, ambiguous matches can be marked as one of the following:

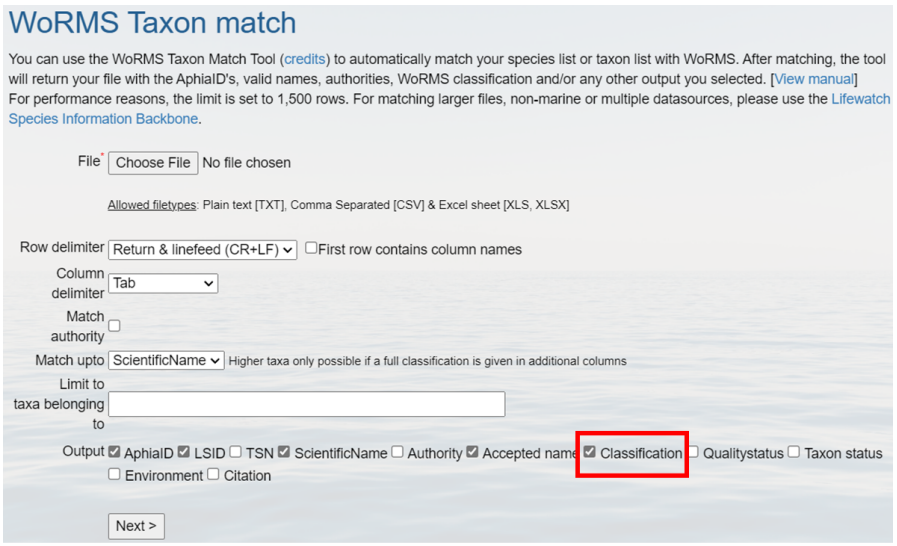


Figure 4.2: WoRMS classification box

- phonetic
- near_1
- near_2
- near_3
- match_quarantine
- match_deleted

See https://www.marinespecies.org/tutorial_taxonmatch.php for definitions of each of these terms.

In each of these cases, WoRMS will try to suggest a species to match your uncertain taxon. Take care to ensure the correct species name is selected. This is especially true for near_2 or near_3 matches. When checking a potential matched name, we recommend referencing the authority and higher taxonomic levels of a given suggestion. For example, if you know the ambiguous species is a sponge, but one of the suggestions is for a mammal, you know that is not the correct name.

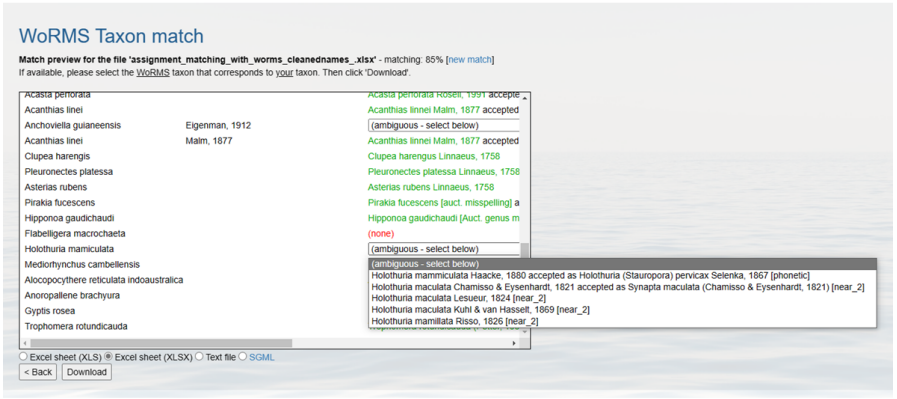


Figure 4.3: Example of choices from an abiguous match

In cases where no match can be found, WoRMS will indicate none. For these cases you should follow these steps:

- Ensure the name was entered correctly and any other information (e.g., authority, year, identification

qualifiers) are included in separate columns, not the same cell as the name.
- Match with LifeWatch or another register (see Step 2 below)
- Check that the species is marine

If a scientific name does not appear in any register, you should contact the original data provider, where possible, to confirm taxonomic spelling, authority, and obtain any original description documents, then attempt to match again. If even after this there are no matches, please contact the WoRMS data management team at info@marinespecies.org to see if the taxon should be added to the WoRMS register.

4.2.1.4 Step 2: Match with other registers

If you do not find a match with WoRMS, you should next check other registers. The LifeWatch taxon match compares your taxon list to multiple taxonomic standards. Matching with multiple registers gives an indication of the correct spelling of a name, regardless of its environment. If a name would not appear in any of the registers, this could indicate a mistake in the scientific name and the name should go back to the provider for additional checking/verification.

Contrary to the WoRMS taxon match, when several matching options are available, the LifeWatch taxon match only mentions "no exact match found, multiple possibilities" instead of listing the available options. If multiple options are available, these should be looked up and matched manually.

Currently, this web service matches the scientific names with the following taxonomic registers:

- World Register of Marine Species – WoRMS
- Catalogue of Life – CoL
- Integrated Taxonomic Information System – ITIS
- Pan-European Species-directories Infrastructure – PESI
- Index Fungorum – IF
- International Plant Names Index – IPNI
- Global Names Index - GNI
- Paleobiology Database - PaleoDB

4.2.1.5 Step 3: Is taxon marine?

The Interim Register of Marine and Non-marine Genera (IRMNG) matching services are available through http://www.irmng.org/, as well as through the LifeWatch taxon match. This service allows you to search for a genus (or other taxonomic rank when you uncheck the "genera" box) to check if it is known to be marine, brackish, freshwater, or terrestrial. You can find this information in the row labeled "Environment". If the taxa is marine, you may have to contact the WoRMS data management team (info@marinespecies.org) to have the taxon added to the WoRMS register (note you may have to provide supporting information confirming taxonomic and marine status).

4.2.2 R packages for taxon matching

If you are familiar with R, you may use the obistools function `match_taxa` to conduct taxon matching for your dataset. There is also a WoRMS package called worrms that has a function called `wm_records_taxamatch` you can use to conduct taxon matching.

The output will be the same as that from the WoRMS tool, so you should check ambiguous matches as described above, confirming with other registers as necessary.

4.2.3 Taxon Match Tools Overview

See the table below for a summary of the different tools available.

Tool	Advantage	Disadvantage
WoRMS taxon match	Accessible online, Does not require coding knowledge	Requires rematch information back to your data
obistools::match_taxa	Produces same output as WoRMS taxon match, Already in R so easier to merge back with data	Requires knowledge of R or python
worrms::wm_records_taxamatch	Outputs all WoRMS matching information	Outputs a tibble for each taxa name specified, Requires knowledge of R or python

4.3 How to format Occurrence tables

If your dataset structure is based on Occurrence core, or has an Occurrence extension (remember that all OBIS data have at least one occurrence record associated, regardless of what organization structure you have chosen), there are several terms that are required in your dataset by OBIS. These required data fields include the following eight terms:

- `occurrenceID`
- `occurrenceStatus`
- `basisOfRecord`
- `scientificName`
- `scientificNameID` (strongly recommended)
- `eventDate` (not required for Occurrence extension, required for Occurrence Core)
- `decimalLatitude` (not required for Occurrence extension)
- `decimalLongitude` (not required for Occurrence extension)

While these are the bare minimum, you should strongly consider adding other terms if you have the corresponding information/data in your dataset or documentation. Other terms you should consider adding are identified by their associated Darwin Core class below. See the term checklist for a more complete list of potential terms for Occurrence table.

- Class Occurrence| DwC: associatedMedia
- Class Occurrence| DwC: associatedReferences
- Class Occurrence| DwC: associatedSequences
- Class Occurrence| DwC: associatedTaxa
- Class Occurrence| DwC: preparations
- Class Occurrence| DwC: recordedBy
- Class Occurrence| DwC: materialSample
- Class Occurrence| DwC: materialSampleID
- Class Record | DwC: bibliographicCitation
- Class Record | DwC: catalogNumber
- Class Record | DwC: collectionCode
- Class Record | DwC: collectionID
- Class Record | DwC: dataGeneralizations
- Class Record | DwC: datasetName
- Class Record | DwC: institutionCode
- Class Record | DwC: modified
- Class Taxon | DwC: kingdom
- Class Taxon | DwC: scientificNameAuthorship
- Class Taxon | DwC: taxonRank
- Class Taxon | DwC: taxonRemarks

Note that any terms related to measurements, either biotic (e.g., sex, lifestage, biomass) or abiotic will be included in extendedMeasurementOrFact table not the Occurrence table.

4.3.0.1 Stepwise Guidance to Format an Occurrence Table (with spreadsheets)

Before proceeding with formatting the Occurrence table, be sure you have completed taxon matching to obtain WoRMS LSIDs for the scientificNameID field.

1. Identify columns in your raw data that match with Occurrence fields
    - Include columns with measurements for now, but they will be moved to an eMoF table(s)
2. Copy these columns to a new sheet named Occurrence (note it is good practice to never make changes to your original datasheet)
3. Create and add occurrenceIDs for each unique occurrence record
4. Add and fill `basisOfRecord` and `occurrenceStatus` fields
5. Ensure your column names map to Darwin Core terms
    - scientificName + scientificNameID

Watch our video tutorial for a demonstration of this procedure:

After formatting your Occurrence Core or Extension table, you can format your extendedMeasurementOrFact table.

4.4 How to format Event tables

If your dataset uses an Event Core structure, data fields included in your dataset should include the following required terms:

- `eventDate`
- `eventID`
- `parentEventID` (if applicable)
- `decimalLatitude`
- `decimalLongitude`

Other terms you should consider adding are grouped by their associated Darwin Core class in the table below. See the term checklist for a more complete list of DwC terms for the Event table.

- Class Event | DwC:parentEventID
- Class Event | DwC:eventRemarks
- Class Event | DwC:eventType
- Class Event | DwC:year
- Class Event | DwC:month
- Class Event | DwC:day
- Class Event | DwC:type
- Class Location | DwC:country
- Class Location | DwC:island
- Class Location | DwC:coordinateUncertaintyInMeters
- Class Location | DwC:countryCode
- Class Location | DwC:footprintWKT
- Class Location | DwC:geodeticDatum
- Class Location | DwC:islandGroup
- Class Location | DwC:locality
- Class Location | DwC:locationAccordingTo
- Class Location | DwC:locationID
- Class Location | DwC:locationRemarks
- Class Location | DwC:maximumDepthInMeters
- Class Location | DwC:minimumDepthInMeters
- Class Location | DwC:stateProvince
- Class Location | DwC:verbatimCoordinates
- Class Location | DwC:verbatimDepth
- Class Location | DwC:waterBody

Terms related to measurements, either biotic (e.g., sex, lifestage) or abiotic will be included in extendedMeasurementOrFact table *not* the Event Core or Occurrence extension table.

4.4.0.1 Stepwise Guidance to Format Event Table (with spreadsheets)

Before proceeding with the below, make sure each record already has an `eventID`.

1. Add and fill the `parentEventID` and `eventRemarks` fields as applicable
2. Identify the hierarchical event structure in your data, if present and create new records for parent Events, filling in any relevant fields
3. Identify all columns in your data that will match with Darwin Core Event fields
   - Include any relevant abiotic measurements (ENV-DATA) related to sampling events (e.g. sampling protocols). We will add these to the eMoF table later
4. Copy these columns to a new sheet and name it Event
5. Delete duplicate data so only unique events are left
6. Ensure dates and time are formatted according to ISO 8601 standards in the `eventDate` field
7. Add any other relevant fields as indicated above
8. Map fields to Darwin Core

Watch the video tutorial of this process below.

After completing the formatting of your Event Core table, you can next format your extendedMeasurementOrFact table. To format the Occurrence extension table, see the Occurrence table section of this manual. Note that there is a difference between how OBIS and GBIF populate fields in parent and child events. In OBIS, child events inherit parent event information. However, if `parentEvent` contains latitude/longitude coordinates, and child events do not, occurrences associated with the child events will have blank latitude/longitude coordinates because these fields are not currently inherited by parent events by GBIF. If you intend for your dataset to be published to GBIF you may consider populating both parent and child events. A discussion of this and the implications can be found here.

4.5 How to format extendedMeasurementOrFact tables

4.5.0.1 What data goes into eMoF

Any data related to abiotic or biotic measurements, including sampling information and protocols should be included in the eMoF table. Measurement data can also go into the MeasurementOrFact extension, however OBIS recommends using the extendedMeasurementOrFact instead, particularly if your data is based on an Event core table.

Required terms for eMoF include:

- `eventID` (this links the record to the Event Core table)
- `occurrenceID` (this links the record to the Occurrence Core or Occurrence Extension table)

Other potential fields are shown in the table below (also listed in the checklist):

- Class Event | DwC:habitat
- Class Event | DwC:sampleSizeUnit
- Class Event | DwC:sampleSizeValue
- Class Event | DwC:samplingEffort
- Class Event | DwC:samplingProtocol
- Class Occurrence | DwC:behavior
- Class Occurrence | DwC:individualCount
- Class Occurrence | DwC:lifeStage
- Class Occurrence | DwC:organismQuantity
- Class Occurrence | DwC:organismQuantityType
- Class Occurrence | DwC:sex
- Class Occurrence | DwC:type
- Class Measurement | DwC:measurementAccuracy
- Class Measurement | DwC:measurementDeterminedBy

- Class Measurement | DwC:measurementDeterminedDate
- Class Measurement | DwC:measurementID
- Class Measurement | DwC:measurementMethod
- Class Measurement | DwC:measurementRemarks
- Class Measurement | DwC:measurementType*
- Class Measurement | DwC:measurementTypeID*
- Class Measurement | DwC:measurementUnit*
- Class Measurement | DwC:measurementUnitID*
- Class Measurement | DwC:measurementValue
- Class Measurement | DwC:measurementValueID

*For `measurementTypeID`, `measurementUnitID`, and `measurementValueID` you must use controlled vocabulary terms. We know choosing the correct vocabulary term can be challenging, so we have provided some guidance on how to select the correct vocabulary. It is strongly recommended to ensure these fields are filled as correctly as possible. Missing or incorrect terms will be documented in the measurementOrFact reports.

4.5.0.2 How to structure eMoF

Structuring data for the eMoF extension may be one of the more confusing extensions in the data formatting process. It may help to think of this extension as the table that contains all information related to any kind of measurement.

Rather than documenting each of your measurements in separate columns (e.g., columns for biomass, abundance, length, gear size, percent cover, etc.), these measurements will be condensed into one column: `measurementValue`. `measurementType` describes what the measurement actually is, for example whether it is an abundance value, length, percent cover, or any other biotic/abiotic measurement. `measurementUnit` is used to indicate the unit of the measurement.

By linking `measurementType` and `measurementValue` with the identifiers `eventID` and/or `occurrenceID`, you can have measurements linked to *one* event (e.g. temperature), measurements link to occurrence records (e.g. length), as well as sampling facts that are linked to events (size, gear, etc.). Information specifically related to how samples were taken will have the measurementTypes: `sampleSizeValue`, `sampleSizeUnit`, `samplingEffort`, and `samplingProtocol`.

4.5.0.3 Stepwise Guidance to Format eMoF Table (in Excel)

1. Create a blank sheet and name it eMoF
2. Add 9 column headers for:
   - `eventID`, `occurrenceID`, `measurementType`, `measurementValue`, `measurementUnit`, `measurementTypeID`, `measurementValueID`, `measurementUnitID`, `measurementRemarks`
3. Copy `eventID` values from your Occurrence table and paste into the `eventID` field in your new, blank eMoF table
   - Repeat for `occurrenceID` from the Occurrence table
4. Copy the first column of measurement values, paste into the `measurementValue` field
   - Fill `measurementType` with the name of the variable (e.g., count, length, etc.)
5. Add unit of measurements where applicable to the `measurementUnit` field
6. For any other measurements related to occurrences, repeat steps 3-5, pasting additional measurements below the preceding ones
   - Be sure to copy and paste the associated occurrenceIDs and/or eventIDs for the additional measurements
7. Fill the fields `measurementTypeID`, `measurementUnitID`, and `measurementValueID` with controlled vocabularies that suit your data (see vocabulary guidelines)
8. Repeat steps 3-7 for any measurements in the Event table

Note the fields sampleSizeValue, samplingEffort, and samplingProtocol from the Occurrence table can be

documented as separate measurements on different rows in the eMoF table. E.g., `measurementType` = samplingProtocol, `measurementValue` = description of protocol. Any values in sampleSizeUnit fields should be placed in the `measurementUnit` field when transferred to the eMoF.

Watch the video tutorial for how to format the eMoF table below.

If you would like to export Event data to the eMoF, see some example R code below. This example was provided by Abby Benson from the OBIS-USA node.

```
library(dplyr)
cruise <- unique(eventCore[c("eventID")]) #create a list of all unique eventIDs from your event table
cruise <- cruise %>% #add sampling information
mutate(measurementType = "Sampling platform name",
measurementValue = "R/V Cruise Id = SR1812",
measurementValueID = "https://doi.org/10.7284/908021",
measurementUnit = "",
measurementTypeID = "http://vocab.nerc.ac.uk/collection/Q01/current/Q0100001/",
measurementUnitID = "",
occurrenceID = "")
```

4.6 DNA derived data

Contents:

4.6.1 Introduction to DNA data

DNA derived data are increasingly being used to document taxon occurrences. This genetic data may come from a sampling event, an individual organism, may be linked to physical material (or not), or may result from DNA detection methods e.g., metabarcoding or qPCR. Thus genetic data may reflect a single organism, or may include information from bulk samples with many individuals. Still, DNA-derived occurrence data of species should be documented as standardized and as reproducible as possible.

To ensure DNA data are useful to the broadest possible community, a community guide entitled Publishing DNA-derived data through biodiversity data platforms was published by GBIF, OBIS, and others. This guide is supported by the DNA derived data extension for Darwin Core, which incorporates MIxS terms into the Darwin Core standard. There are 5 categories for which genetic data could fall into:

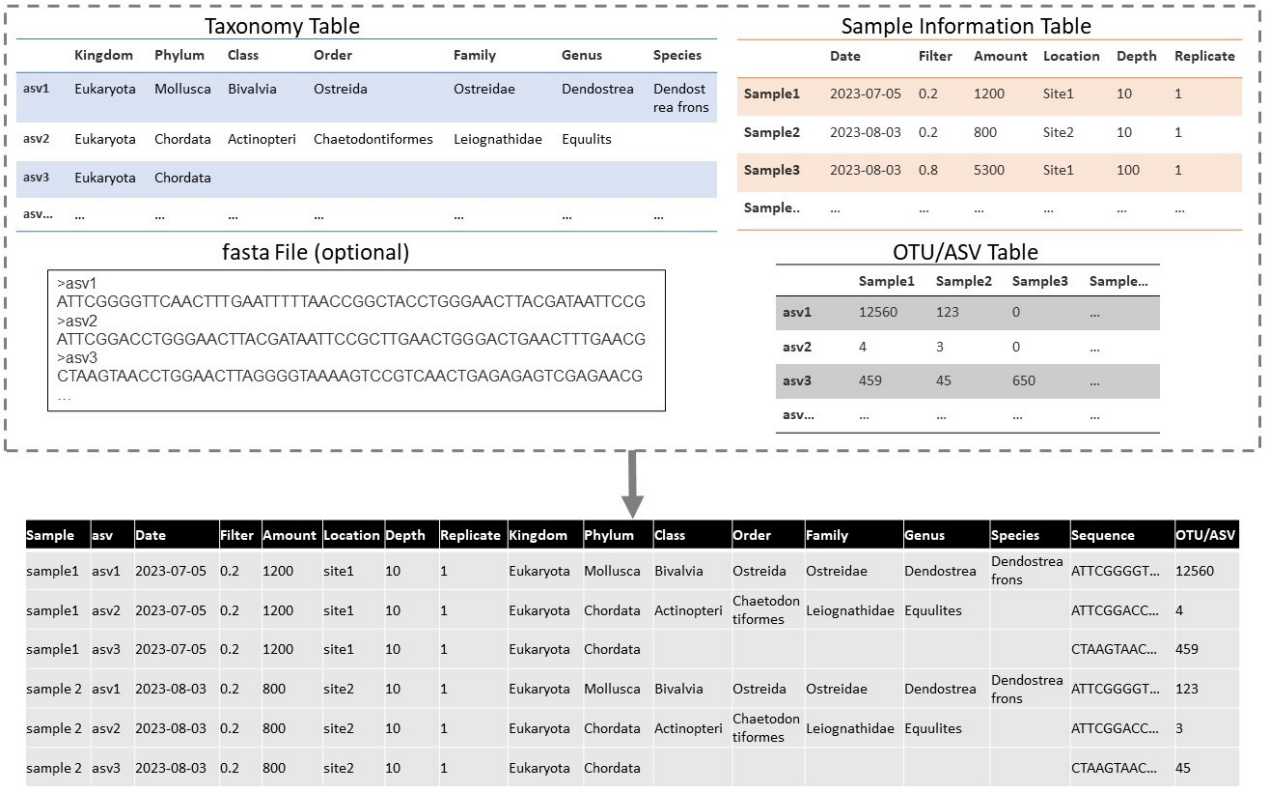
1. DNA-derived occurrences
2. Enriched occurrences
3. Targeted species detection
4. Name references
5. Metadata only

For a guide and decision tree on determining which category your data falls into, see the Data packaging and mapping section of the GBIF guide. Refer to the examples below for use case examples of eDNA and DNA derived data (Category 1).

> Currently, genetic data **must** be published with Occurrence core, not Event core. eDNA and DNA derived data are then linked to the Occurrence core data table with the use of `occurrenceID` and/or `eventID`. See below for further guidance on compiling genetic data. A new data model is being developed by GBIF and the OBIS community that may change this, however as it is not implemented yet, we focus on the current Darwin Core recommendations here.

To format datasets, you will need to have information on the sequence and possible taxonomy for each occurrence record associated with a DNA sample. Genetic data is often recorded in multiple different files, and this might be the type of format received from data providers. Important data tables can include: an OTU-table, a taxonomy table, a sample information table, and a .fasta file with sequences. The OTU-table is a sequence by sample table, which records the quantity of each unique sequence found in each sample. Sequences are usually referred to by an ID, which is unique only in the dataset (e.g. asv1, asv2, asv3 . . .). The taxonomy table is a sequence by taxonomy table, which records the taxonomy linked to each unique sequence, as defined by the annotation method. The sample information table records the metadata of each sample (e.g. location, time, and collection method). Finally the .fasta file records the actual DNA sequence that is linked to each sequence id.

Although this data is in multiple files, each unique sequence by sample combination is considered **one** occurrence. Therefore the data from these tables will need to be formatted to the "long format", including a row for each sequence in each sample. See the figure below for a demonstration of how this can be done.



Sample	asv	Date	Filter	Amount	Location	Depth	Replicate	Kingdom	Phylum	Class	Order	Family	Genus	Species	Sequence	OTU/ASV
sample1	asv1	2023-07-05	0.2	1200	site1	10	1	Eukaryota	Mollusca	Bivalvia	Ostreida	Ostreidae	Dendostrea	Dendostrea frons	ATTCGGGGT...	12560
sample1	asv2	2023-07-05	0.2	1200	site1	10	1	Eukaryota	Chordata	Actinopteri	Chaetodontiformes	Leiognathidae	Equulites		ATTCGGACC...	4
sample1	asv3	2023-07-05	0.2	1200	site1	10	1	Eukaryota	Chordata						CTAAGTAAC...	459
sample 2	asv1	2023-08-03	0.2	800	site2	10	1	Eukaryota	Mollusca	Bivalvia	Ostreida	Ostreidae	Dendostrea	Dendostrea frons	ATTCGGGGT...	123
sample 2	asv2	2023-08-03	0.2	800	site2	10	1	Eukaryota	Chordata	Actinopteri	Chaetodontiformes	Leiognathidae	Equulites		ATTCGGACC...	3
sample 2	asv3	2023-08-03	0.2	800	site2	10	1	Eukaryota	Chordata						CTAAGTAAC...	45

Figure 4.4: Combining multiple DNA data tables into one "long format" table

Doing this will help you when following the guidelines below.

4.6.2 How to find genetic data in OBIS

To find genetic data in OBIS we recommend using the R package robis, using the `occurrence` function. You must set `extensions` and/or `hasextensions` to "DNADerivedData" to ensure extension records are included in the results. `hasextensions` will exclude any occurrence that does not have the specified extension, in our case, DNADerivedData. The `extensions` parameter specifies which extensions to include. To obtain the DNA data, you have to extract the information from the extension using the `unnest_extension()` function. You can specify as many fields from the Occurrence table to be included, and pass them to the `fields` parameter. See the code below for an example. See also this vignette for a more detailed example, including how you can work further with these sequences in R.

```
dna_occ<-occurrence("Dinophyceae",hasextensions="DNADerivedData", extensions="DNADerivedData")
dnaseqs <-unnest_extension(dna_datasets, "DNADerivedData", fields = c("id", "phylum", "class", "family",
dnaseqs["DNA_sequence"]
# A tibble: 706 × 1
   DNA_sequence
   <chr>
 1 AGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAACGCTCGTAGTCGGATTTCGGGGCGGGCCGACCGGTCTGCCGATGGGTATGCACTGGCCGGCGC
 2 AGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAACGCTCGTAGTCGGATTTCGGGGCGGGCCGACCGGTCTGCCGATGGGTATGCACTGGCCGGCGC
 3 GCTCCTACCGATTGAATGATCCGGTGAGGCCCCCGGACTGCGGCGCCGCAGCTGGTTCTCCAGCCGCGACGCCGCGGGAAGCTGTCCGAACCTTATCATTT
 4 AGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAACGCTCGTAGTCGGATTTCGGGGCGGGCCGACCGGTCTGCCGATGGGTATGCACTGGCCGGCGC
 5 AGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAACGCTCGTAGTCGGATTTCGGGGCGGGCCGACCGGTCTGCCGATGGGTATGCACTGGCCGGCGC
 6 GCTCCTACCGATTGAATGATCCGGTGAGGCCCCCGGACTGCGGCGCCGCAGCTGGTTCTCCAGCCGCGACGCCGCGGGAAGCTGTCCGAACCTTATCATTT
 7 AGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAACGCTCGTAGTCGGATTTCGGGGCGGGCCGACCGGTCTGCCGATGGGTATGCACTGGCCGGCGC
 8 AGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAACGCTCGTAGTCGGATTTCGGGGCGGGCCGACCGGTCTGCCGATGGGTATGCACTGGCCGGCGC
 9 GCTCCTACCGATTGAATGATCCGGTGAGGCCCCCGGACTGCGGCGCCGCAGCTGGTTCTCCAGCCGCGACGCCGCGGGAAGCTGTCCGAACCTTATCATTT
10 AGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAACGCTCGTAGTCGGATTTCGGGGCGGGCCGACCGGTCTGCCGATGGGTATGCACTGGCCGGCGC
# ... with 696 more rows
```

Additionally, a prototype Sequence Search tool is in development that allows you to search for sequences or related sequences in OBIS. Note that the tool is not always up to date so use caution until the prototype is fully developed. To use the tool:

1. Copy your sequence in the provided box (an example sequence is provided for testing)
2. Press the Search button
3. View results below
4. You can also change the Minimum Alignment Score slider in the map view to see location of sequences

The search result will show you taxonomic information for species sequences that align to your sequence, the alignment score, and a link to the respective datasets.

OBIS Mapper

You can use the OBIS Mapper to obtain records that include the DNADerivedData extension by addding a filter for the extension when using the tool. To do this:

1. From the OBIS Mapper, navigate to the Criteria tab (the plus (+) sign)
2. Open the Extensions dropdown section
3. Check the box for DNADerivedData
4. Add any other filters you want, e.g. taxonomic, then click save to create the layer
5. Switch to the Layers tab
6. Download the data from the layer by clicking the green button (see Data Access for more on using the OBIS Mapper)

4.6.3 Guidelines for compiling genetic data: eDNA and metabarcoding datasets

As mentioned above, you will need to have information on the taxonomy and sequences for each occurrence record associated with a DNA sample. You should first fill in the Occurrence core table, and then complete the DNA Derived Data extension (as well as the eMoF extension, if applicable, for any measurements taken).

Occurrence core table In addition to the usual required terms for Occurrence datasets, you should consider the following additional terms:

- Class Occurrence | DwC: organismQuantity
- Class Occurrence | DwC: OrganismQuantityType
- Class Occurrence| DwC: associatedSequences
- Class Event | DwC: sampleSizeValue
- Class Event | DwC: sampleSizeUnit
- Class Event | DwC: samplingProtocol
- Class Identification | DwC: identificationRemarks
- Class Identification | DwC: identificationReferences
- Class Identification | DwC: verbatimIdentification
- Class Taxon | DwC: taxonConceptID
- Class Material Sample | DwC:materialSampleID

For `organismQuantity` and `sampleSizeValue` in eDNA datasets, the quantities recorded with sequencing studies always represent relative abundance to the total reads in the sample, and cannot be directly compared across samples. This is due to the nature of the sample processing protocol and the amplification of DNA with PCR, which biases the original quantities. In `organismQuantity`, record the amount of a unique sequence in a specific sample (i.e. 33 reads). In `sampleSizeValue`, record the total number of all reads in that specific sample (i.e. 15310 reads). This information will allow people accessing the data to calculate the relative abundance of that sequence in the sample. The fields `organismQuantityType`, and `sampleSizeUnit`, should be populated with "DNA sequence reads", as it is of high importance that sequence abundances are not confused with organism abundances recorded by traditional methods. The abundance information can usually be found in the "OTU-table".

`associatedSequences` should contain a link to the "raw" sequences deposited in a public database or list of identifiers for the genetic sequence associated with the occurrence record (e.g. GenBank). The actual sequence of the occurrence will be documented in the DNA Derived Data extension.

`identificationRemarks` should be used to record information on how the taxonomic information of the occurrence was reached against which reference database, and, if possible, with which confidence. For example "RDP annotation confidence: 0.96, against reference database: GTDB". This information should be recorded in the bioinformatic protocol of the study. Note: this information will also be recorded in the DNA derived extension in the fields `otu_seq_comp_appr` and `otu_db`.

`identificationReferences` should include a link to the bioinformatic pipeline or publication where the identification process is explained in detail.

`taxonConceptID` should include the taxonomic ID of the sequence (non-Linnean). Often genetic sequences can be assigned an ID linked to a reference database that is not a linnean name. These taxonomic names or IDs from other taxonomic databases (like the NCBI taxonomic database) can be recorded in this field. For example: NCBI:txid9771. The name linked to this ID can then be recorded in the field `verbatimIdentification`.

`samplingProtocol` can contain free-text that briefly describes the methods used to obtain the sample, or a link to a protocol that is recorded elsewhere.

DNA Derived Data extension The DNADerivedData extension is meant to capture information related to the sampled DNA, including sampling, processing, and other bioinformatic methods. The following (free-text) terms are required or highly recommended for eDNA and metabarcoding datasets. Note that some terms will be different for qPCR data (see below)

- DNA Derived | DwC: DNA_sequence
- DNA Derived | DwC: sop
- DNA Derived | DwC: target_gene
- DNA Derived | DwC: target_subfragment
- DNA Derived | DwC: pcr_primer_forward
- DNA Derived | DwC: pcr_primer_reverse
- DNA Derived | DwC: pcr_primer_name_forward
- DNA Derived | DwC: pcr_primer_name_reverse
- DNA Derived | DwC: pcr_primer_reference
- DNA Derived | DwC: Pcr_cond
- DNA Derived | DwC: annealingTemp
- DNA Derived | DwC: annealinTempUnit
- DNA Derived | DwC: ampliconSize
- DNA Derived | DwC: env_broad_scale
- DNA Derived | DwC: env_local_scale
- DNA Derived | DwC: env_medium
- DNA Derived | DwC: lib_layout
- DNA Derived | DwC: seq_meth
- DNA Derived | DwC: otu_class_appr
- DNA Derived | DwC: otu_seq_comp_appr
- DNA Derived | DwC: otu_db

For a complete list of terms you can map to, see the DwC DNA Derived Data extension page. See the examples below for use case examples. The Marine Biological Data Mobilization Workshop also has a tutorial for this type of data.

`DNA_sequence` is the most important field, where the ASV/OTU sequence will be recorded. This field can then be searched with sequence alignment methods to, for example, find closely related sequences recorded in other studies, and will allow very powerful data comparison and analysis in the future. It will also make your sequence available in the OBIS sequence search tool.

The remaining metadata fields will help the person accessing the data to filter data of interest (e.g. specific genetic region with `target_gene`, `target_subfragment`, or `pcr_primer` fields), link to the public sequence databases with the MixS specific fields (e.g. `env_` fields), and evaluate the reliability of the sequence annotation method (e.g. `otu_` fields).

Environmental systems are described in the two fields `env_broad_scale` and `env_local_scale` and it is recommended to use Environment Ontology (ENVO)'s biome classes to describe the environmental system from which the sample was extracted. Like other identifiers, provide the exact ENVO reference identifier. `env_broad_scale` provides a coarse resolution for which environment your sample came from. Likely this will be marine biome (ENVO:00000447) for OBIS data. For local scale, identify the specific environment your sample was obtained from (e.g., coastal water, benthic zone, etc.).

When data tables are formatted and you are ready to publish it on the IPT, it will follow the same process for publishing on an IPT. You will upload your source files, and add the Occurrence core Darwin Core mappings, and then the DNA Derived Data Darwin Core mappings. However the extension must first be installed by the IPT administrator (often the node manager). Once the extension is installed, you can add the Darwin Core DNA Derived Data mapping for that file.

4.6.3.1 eDNA and DNA derived data example

The following example use cases draw on both the DNA-derived data guide and the DNA derived data extension to illustrate how to incorporate a DNA derived data extension file into a Darwin Core archive. Note: for the purposes of this section, only required Occurrence core terms are shown, in addition to all eDNA & DNA specific terms. For additional Occurrence core terms, refer to Occurrence.

4.6.3.1.1 eDNA data from Monterey Bay, California The data for this example is from the use case "18S Monterey Bay Time Series: an eDNA data set from Monterey Bay, California, including years 2006, 2013 - 2016'. The data from this study originate from marine filtered seawater samples that have undergone metabarcoding of the 18S V9 region.

Occurrence core:

We can populate the Occurrence core with all the required and highly recommended fields, as well as considering the eDNA and DNA specific fields. The Occurrence core contain the taxonomic identification of each ASV observed; its number of reads, as well as relevant metadata including the sample collection location, references for the identification procedure, and links to archived sequences.

`OccurrenceID` and `basisOfRecord` are some of the required Occurrence core terms, in addition to the highly recommended fields of `organismQuantity` and `organismQuantityType`. A selection of samples from this plate were included in another publication (Djurhuus et al., 2020),which is recorded in `identificationReferences` along with the GitHub repository where the data can be found.

occurrenceID	basisOfRecord	organismQuantity	OrganismQuantityType	associatedSequences
11216c01_12_edna_1_S_occ1	MaterialSample	19312	DNA sequence reads	NCBI BioProject acc. nr. PRJNA433203
11216c01_12_edna_2_S_occ1	MaterialSample	16491	DNA sequence reads	NCBI BioProject acc. nr. PRJNA433203
11216c01_12_edna_3_S_occ1	MaterialSample	21670	DNA sequence reads	NCBI BioProject acc. nr. PRJNA433203

sampleSizeValue	sampleSizeUnit	identificationReferences	identificationRemarks
147220	DNA sequence reads	GitHub repository Djurhuus et al. 2020	unassigned, Genbank nr Release 221 September 20 2017
121419	DNA sequence reads	GitHub repository Djurhuus et al. 2020	unassigned, Genbank nr Release 221 September 20 2017
161525	DNA sequence reads	GitHub repository Djurhuus et al. 2020	unassigned, Genbank nr Release 221 September 20 2017

DNA Derived Data extension:

Next, we can create the **DNA Derived Data extension** which will be connected to the Occurrence core with the use of `occurrenceID`. This extension contains the DNA sequences and relevant DNA metadata, including sequencing procedures, primers used and SOP's. The recommended use of ENVO's biome classes were applied to describe the environmental system from which the sample was extracted. The samples were collected by CTD rosette and filtered by a peristaltic pump system. Illumina MiSeq metabarcoding was applied for the target_gene 18S and the target_subfragment, V9 region. URL's are provided for the protocols followed for nucleic acids extraction and amplification.

For a detailed description of the steps taken to process the data, including algorithms used, see the original publication. Adding Operational Taxonomic Unit (OTU) related data are highly recommended and should be as complete as possible, for example:

occurrenceID	env-broad_scale	env_local_scale	env_medium
11216c01_12_edna_1_S_occ1	marine biome (ENVO:00000447)	coastal water (ENVO:00001250)	waterborne particulate matter (ENVO:01000436)
11216c01_12_edna_2_S_occ1	marine biome (ENVO:00000447)	coastal water (ENVO:00001250)	waterborne particulate matter (ENVO:01000436)
11216c01_12_edna_3_S_occ1	marine biome (ENVO:00000447)	coastal water (ENVO:00001250)	waterborne particulate matter (ENVO:01000436)

samp_vol_we_dna_ext	nucl_acid_ext	nucl_acid_amp	lib_layout	target_gene
1000ml	dx.doi.org/10.17504/protocols.io.xjufknw	dx.doi.org/10.17504/protocols.io.n2vdge6	paired	18S
1000ml	dx.doi.org/10.17504/protocols.io.xjufknw	dx.doi.org/10.17504/protocols.io.n2vdge6	paired	18S
1000ml	dx.doi.org/10.17504/protocols.io.xjufknw	dx.doi.org/10.17504/protocols.io.n2vdge6	paired	18S

target_subfragment	seq_meth	otu_class_appr	otu_seq_comp_appr
V9	Illumina MiSeq 2x250	dada2;1.14.0;ASV	blast;2.9.0+;80% identity;e-value cutoff: x MEGAN6;6.18.5;bitscore: 100 :2%
V9	Illumina MiSeq 2x250	dada2;1.14.0;ASV	blast;2.9.0+;80% identity;e-value cutoff: x MEGAN6;6.18.5;bitscore: 100 :2%
V9	Illumina MiSeq 2x250	dada2;1.14.0;ASV	blast;2.9.0+;80% identity;e-value cutoff: x MEGAN6;6.18.5;bitscore: 100 :2%

otu_db	sop	DNA_sequence
Genbank nr;221	dx.doi.org/10.17504/protocols.io.xjufknw or GitHub repository	GCTACTACCGATT...
Genbank nr;221	dx.doi.org/10.17504/protocols.io.xjufknw or GitHub repository	GCTACTACCGATT...
Genbank nr;221	dx.doi.org/10.17504/protocols.io.xjufknw or GitHub repository	GCTACTACCGATT...

pcr_primer_forward	pcr_primer_reverse	pcr_primer_name_forward	pcr_primer_name_reverse	pcr_primer_reference
GTACACACCGCCCGTC	TGATCCTTCTGCAGGTTCACCTAC	1391f	EukBr	Amaral-Zettler et al. 2009
GTACACACCGCCCGTC	TGATCCTTCTGCAGGTTCACCTAC	1391f	EukBr	Amaral-Zettler et al. 2009
GTACACACCGCCCGTC	TGATCCTTCTGCAGGTTCACCTAC	1391f	EukBr	Amaral-Zettler et al. 2009

4.6.3.1.2 16S rRNA gene metabarcoding data of Pico- to Mesoplankton

DNA derived datasets can also include an extendedMeasurementsOrFact (eMoF) extension file, in addition to the Occurrence and DNA derived extensions. In this example, environmental measurements were provided in an eMoF file, in addition to the DNA derived data and occurrence data. Here we show how to incorporate such measurements in the extensions.

In the publication "Diversity of Pico- to Mesoplankton along the 2000 km Salinity Gradient of the Baltic Sea", a dataset with 16S rRNA gene metabarcoding data of surface water microbial communities was created from 21 off-shore stations, following a transect from Kattegat to the Gulf of Bothnia in the Baltic Sea. The full dataset entitled "Diversity of Pico- to Mesoplankton along the 2000 km Salinity Gradient of the Baltic Sea (Hu et al. 2016) is available from GBIF.

Occurrence core:

The Occurrence core contain information about the organisms in the sample including the taxonomy and quantity of organisms detected, the collection location, references for the identification procedure, and links to the sequences generated.

Important note: even though this dataset uses OTU identifiers for taxonomy (therefore not including `scientificNameID`) OBIS still recommends using `scientificNameID`.

basisOfRecord	occurrenceID	eventID	eventDate
MaterialSample	SBDI-ASV-3:16S_1:919a2aa9d306e4cf3fa9ca02a2aa5730	SBDI-ASV-3:16S_1	2013-07-13 07:08:00
MaterialSample	SBDI-ASV-3:16S_1:43e088977eba5732bfa45e20b1d8cdd2	SBDI-ASV-3:16S_1	2013-07-13 07:08:00
MaterialSample	SBDI-ASV-3:16S_1:887bc7033b46d960e893caceb711700b	SBDI-ASV-3:16S_1	2013-07-13 07:08:00

organismQuantity	organismQuantityType	sampleSizeValue	sampleSizeUnit
2235	DNA sequence reads	12393	DNA sequence reads
795	DNA sequence reads	12393	DNA sequence reads
40	DNA sequence reads	12393	DNA sequence reads

samplingProtocol	associatedSequences	identificationReferences	identificationRemarks
200–500 mL seawater were filtered onto 0.22 μm pore-size mixed cellulose ester membrane filters; [https://doi.org/10.3389/fmicb.2016.00679]	[https://www.ebi.ac.uk/ena/browser/view/ERR1202034]	[https://docs.biodiversitydata.se/analyse-data/molecular-tools/#taxonomy-annotation]	DADA2:assignTaxonomy:addSpecies annotation against sbdi-gtdb=R06-RS202-1; confidence at lowest specified (ASV portal) taxon: 0.5
200–500 mL seawater were filtered onto 0.22 μm pore-size mixed cellulose ester membrane filters; [https://doi.org/10.3389/fmicb.2016.00679]	[https://www.ebi.ac.uk/ena/browser/view/ERR1202034]	[https://docs.biodiversitydata.se/analyse-data/molecular-tools/#taxonomy-annotation]	DADA2:assignTaxonomy:addSpecies annotation against sbdi-gtdb=R06-RS202-1; confidence at lowest specified (ASV portal) taxon: 0.56
200–500 mL seawater were filtered onto 0.22 μm pore-size mixed cellulose ester membrane filters; [https://doi.org/10.3389/fmicb.2016.00679]	[https://www.ebi.ac.uk/ena/browser/view/ERR1202034]	[https://docs.biodiversitydata.se/analyse-data/molecular-tools/#taxonomy-annotation]	DADA2:assignTaxonomy:addSpecies annotation against sbdi-gtdb=R06-RS202-1; confidence at lowest specified (ASV portal) taxon: 0.99

decimalLatitude	decimalLongitude	taxonID	scientificName
55.185	13.791	ASV:919a2aa9d306e4cf3fa9ca02a2aa5730	UBA6821
55.185	13.791	ASV:43e088977eba5732bfa45e20b1d8cdd2	Chthoniobacterales
55.185	13.791	ASV:887bc7033b46d960e893caceb711700b	BACL27 sp014190055

kingdom	phylum	class	order	family	genus
Bacteria	Verrucomicrobiota	Verrucomicrobiae	Chthoniobacterales	UBA6821	UBA6821
Bacteria	Verrucomicrobiota	Verrucomicrobiae	Chthoniobacterales	NA	NA
Bacteria	Actinobacteriota	Acidimicrobiia	Acidimicrobiales	Ilumatobacteraceae	BACL27

DNA Derived Data extension:

The DNA Derived Data extension for metabarcoding data contains the DNA sequences and relevant DNA metadata, primers and procedures. This example table contains the highly recommended and recommended fields as populated with the example dataset data. For this dataset, authors additionally provided measurements of of water sample temperature and salinity, which are provided in an **extendedMeasurementOrFact** extension file:

id	env_broad_scale	env_local_scale	env_medium
SBDI-ASV-3:16S_1:919a2aa9d306e4cf3fa9ca02a2aa5730	aquatic biome [ENVO_00002030]	marine biome [ENVO_00000447]	brackish water [ENVO_00002019]
SBDI-ASV-3:16S_1:43e088977eba5732bfa45e20b1d8cdd2	aquatic biome [ENVO_00002030]	marine biome [ENVO_00000447]	brackish water [ENVO_00002019]
SBDI-ASV-3:16S_1:887bc7033b46d960e893caceb711700b	aquatic biome [ENVO_00002030]	marine biome [ENVO_00000447]	brackish water [ENVO_00002019]

lib_layout	target_gene	target_subfragment	seq_meth	sop
paired	16S rRNA	V3-V4	Illumina MiSeq	https://nf-co.re/ampliseq
paired	16S rRNA	V3-V4	Illumina MiSeq	https://nf-co.re/ampliseq
paired	16S rRNA	V3-V4	Illumina MiSeq	https://nf-co.re/ampliseq

pcr_primer_forward	pcr_primer_reverse	pcr_primer_name_forward	pcr_primer_name_reverse	DNA_sequence
CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	341	805R	TCGAGAATTTTTCACAATG...
CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	341	805R	TCGAGAATTTTTCACAATG...
CCTACGGGNGGCWGCAG	GACTACHVGGGTATCTAATCC	341	805R	TGGGGAATCTTGCGCAATG...

extendedMeasurementOrFact (eMoF) extension:

measurementID	occurrenceID	measurementType	measurementValue	measurementUnit
SBDI-ASV-3:16S_1:temperature	SBDI-ASV-3:16S_1:919a2aa9d306e4cf3fa9ca02a2aa5730	temperature	16.9	°C
SBDI-ASV-3:16S_1:salinity	SBDI-ASV-3:16S_1:919a2aa9d306e4cf3fa9ca02a2aa5730	salinity	7.25	psu
SBDI-ASV-3:16S_1:temperature	SBDI-ASV-3:16S_1:1ead98754d34073a4606f7ff1e94126e	temperature	16.9	°C

4.6.4 Unknown sequences

It is important to understand the significance of unknown and uncharacterized sequences in genetic studies. Sequences are given taxonomic names based on comparisons to a reference database. The reference databases contain sequences that have been submitted with a name. Ideally, the reference database is a collection of sequences that are derived from vouchered, morphologically identified specimens. Notably, this is frequently often not the case and sequences can also have erratic annotations. Furthermore, only a small portion of species have sequences in reference databases. Due to this reason, typically many sequences in any given study will remain uncharacterized. This is especially the case for tropical regions with high biodiversity. By also recording all sequences, including uncharacterized sequences, we make sure that the information is not lost, even if the annotation is currently incorrect or missing. These uncharacterized sequences can then still be compared to other studies, and can be given a taxonomic name as more specimens are sequenced and added to the reference databases.

For unknown sequences it is required to populate the `scientificName` field with "Incertae sedis", or the lowest taxonomic information if available. For example, if it is only known which Class a sequence belongs to, populate `scientificName` with the associated Class name. Similarly, `scientificNameID` should be populated with the WoRMS LSID for the name given to `scientificName`. For records recorded as Incertae sedis,

`scientificNameID` should be populated with urn:lsid:marinespecies.org:taxname:12. We recommend also populating `verbatimIdentification` with the name that was originally documented (e.g. phototrophic eukaryote).

4.6.5 Guidelines for compiling genetic data: qPCR

Compiling qPCR data is a little bit different than compiling eDNA or metabarcoding data. One of the main differences is that there are no sequences recorded in the `DNA_sequence` field of the DNA derived data extension. Instead, occurrences are based on detections made using species-specific primers and either qPCR (Quantitative Polymerase Chain Reaction) or ddPCR (Droplet-Digital Polymerase Chain Reaction), no sequencing is done.

It is very important to document the methods used for this type of data because the results can be sensitive to the specificity of the primers/assays used. Therefore documenting as much detail on the methodologies is important to ensure data interpretability.

Occurrence core table In addition to the Occurrence core terms, it is strongly recommend to including the following terms for qPCR data:

- Class Occurrence | DwC:recordedBy
- Class Occurrence | DwC: organismQuantity
- Class Occurrence | DwC: organismQuantityType
- Class Event | DwC: sampleSizeValue
- Class Event | DwC: sampleSizeUnit
- Class Event | DwC: samplingProtocol
- Class Material Sample | DwC:materialSampleID

For ddPCR, `organismQuantity` refers to the number of positive droplets/chambers in the sample, and `organismQuantityType` is the partition type (e.g., ddPCR droplets, dPCR chambers). `sampleSizeValue` will be populated with the number of accepted partitions, e.g. meaning accepted droplets in ddPCR or chambers in dPCR. `sampleSizeUnit` is the partition type, which should be the same as `organismQuantityType`. All four of these fields are particularly important to include for ddPCR data.

For qPCR, these fields can be used for recording e.g. the number of copies that were calculated for the target gene in the sample. In this case `organismQuantityType` needs to contain the exact type of the measurement reported in the results. The field accepts any string, but the best practice would be to add a URI pointing to a vocabulary, as is done in the extendedMeasurementOrFact extension. The terms `sampleSizeValue` and `sampleSizeUnit` would not be used in this case.

`materialSampleID` should contain an identifier for the MaterialSample (i.e. occurrence record), rather than a digital record of the material sample. If an ID was obtained from a nucleotide archive, use the associated biosample ID. Otherwise, construct a persistent unique identifier from a combination of elements in the data that will make the `materialSampleID` globally unique, similar to eventIDs and occurrenceIDs.

`recordedBy` can be populated with the names of the people, groups, or organizations responsible for recording the original Occurrence. You can use a concatenated list for multiple names, by separating values with a vertical bar (' | ').

DNA Derived Data extension For qPCR datasets, it is strongly recommended to document as much detail as possible in this extension, particularly details about the PCR primers used and the target gene. We recommend you include the following terms, where relevant. For term definitions see DNA derived data extension.

Terms related to the sampling event:

- DNA Derived | DwC: env_broad_scale
- DNA Derived | DwC: env_local_scale
- DNA Derived | DwC: env_medium
- DNA Derived | DwC: samp_collect_device

- DNA Derived | DwC: samp_mat_process
- DNA Derived | DwC: samp_size
- DNA Derived | DwC: size_frac

Terms related to DNA and PCR methods:

- DNA Derived | DwC: sop
- DNA Derived | DwC: concentration
- DNA Derived | DwC: concentrationUnit
- DNA Derived | DwC: methodDeterminationConcentrationAndRatios
- DNA Derived | DwC: contaminationAssessment
- DNA Derived | DwC: target_gene
- DNA Derived | DwC: target_subfragment
- DNA Derived | DwC: ampliconSize
- DNA Derived | DwC: amplificationReactionVolume
- DNA Derived | DwC: amplificationReactionVolumeUnit
- DNA Derived | DwC: baselineValue
- DNA Derived | DwC: automaticBaselineValue
- DNA Derived | DwC: automaticThresholdQuantificationCycle
- DNA Derived | DwC: thresholdQuantificationCycle
- DNA Derived | DwC: pcr_analysis_software
- DNA Derived | DwC: pcr_primer_forward
- DNA Derived | DwC: pcr_primer_reverse
- DNA Derived | DwC: pcr_primer_name_forward
- DNA Derived | DwC: pcr_primer_name_reverse
- DNA Derived | DwC: pcr_primer_reference
- DNA Derived | DwC: pcr_cond
- DNA Derived | DwC: pcr_primer_lod
- DNA Derived | DwC: pcr_primer_loq
- DNA Derived | DwC: annealingTemp
- DNA Derived | DwC: annealinTempUnit
- DNA Derived | DwC: probeQuencher
- DNA Derived | DwC: probeReporter
- DNA Derived | DwC: quantificationCycle
- DNA Derived | DwC: ratioOfAbsorbance260_230
- DNA Derived | DwC: ratioOfAbsorbance260_280

There are many specialized qPCR terms that are possible to add to the dataset. The terms `concentration`, `concentrationUnit`, `ratioOfAbsorbance260_230`, `ratioOfAbsorbance260_280`, and `methodDeterminationConcentrationAndRatios` are related to the original DNA sample before qPCR analysis, and can be useful for evaluating the prevalence of the marker in the sample as well as the purity of the DNA for any indication of PCR inhibition.

As with the metabarcoding dataset, the details of the PCR conditions and primers can be recorded in the multiple `pcr_` terms as well as `target_` terms, and `amplificationReactionVolume` and `amplificationReactionVolumeUnit`.

The main terms that are important for the quantification information and are different from the metabarcoding dataset are `baselineValue`, `thresholdQuantificationCycle` and `quantificationCycle`. The terms `pcr_primer_lod`, `pcr_primer_loq`, `probeQuencher`, `probeReporter` are additional terms specific for qPCR assays. The `baselineValue` indicates the number of cycles below which the signal is considered only background noise. The `quantificationCycle` is the most important and indicates at which cycle the particular sample crossed the detection threshold, this will be different for each sample. It is recommended to record this information, but not all of this may be easily available.

4.6.6 OBIS Bioinformatics Pipline

OBIS recognizes the vast amount of data generated from marine DNA sampling, especially from eDNA sequencing. Thus we have been developing a bioinformatics pipeline to facilitate publication of this data into OBIS. The pipeline was initially developed for the PacMAN project (Pacific Islands Marine Bioinvasions Alert Network).

Broadly speaking, it creates a framework that receives raw sequence data from eDNA samples, cleans, aligns, classifies sequences, and finally outputs a DwC-compatible table. The pipeline is currently under development and for now only accepts CO1 data. It will be extended to include other genetic markers in the future. More details about the PacMAN pipeline can be found on its associated GitHub repository. Once fully online, we will provide guidelines on how to use the pipeline.

OBIS is developing guidelines and pipelines to accept other data types, such as:

- Acoustic
- Imaging
- Tracking
- Habitat

4.7 Choosing vocabularies for your dataset

Content

4.7.1 Map data fields to Darwin Core

There are many possible ways of setting up your datasheets, and if you are new to OBIS you likely did not use controlled Darwin Core (DwC) or BODC vocabulary before samples were collected. In mapping your data fields to DwC we recommend documenting your choices so you have a reference to go back to should the need arise. In such a document you should take notes on the choices you made, as well as any actions you had to take (e.g. separate one column into many, convert dates or coordinates, etc.).

For example, a DwC mapping reference table could look like the following:

Verbatim field name	Mapped DwC term	Actions taken	Notes
date	eventDate	convert dates to ISO	
coordinates	decimalLongitude, decimalLatitude	convert ddmmss to decimal degrees, separated one column into 2 for longitude and latitude	put original coordinates into verbatimCoordinates

In order to help you map your data to DwC terms, we have provided the table below which outlines some common data fields, their associated Darwin Core vocabulary, and which data table the field is likely to go in:

Common Raw Terms	DwC Field	Data table
Date, Time	eventDate	Event, Occurrence
Species, g_s, taxa	scientificName	Occurrence
Any biotic/abiotic measurements*	measurementType, measurementValue, measurementUnit*	eMoF
Depth	maximumDepthInMeters or minimumDepthInMeters	Event, Occurrence
Lat/Latitude, Lon/Long/Longitude, dd	decimalLatitude, decimalLongitude	Event, Occurrence
Sampling method	samplingProtocol	eMoF
Sample size, N, #, No.	sampleSizeValue	eMoF
Location	locality	Event
Presence, absence	occurrenceStatus	Occurrence
Type of record/ specimen	basisofRecord	Occurrence

Common Raw Terms	DwC Field	Data table
Person/ people that recorded the original Occurrence	recordedBy	Occurrence
OrcID of person/ people that recorded the original Occurrence	recordedByID	Occurrence
Person/ people that identified the organism	identifiedBy	Occurrence
OrcID of person/ people that identified the organism	identifiedByID	Occurrence
Data collector, data creator	recordedBy	Event, Occurrence
Taxonomist, identifier	identifiedBy	Occurrence
Record number, sample number, observation number	occurrenceID (either ID or incorporated into ID)	Occurrence

Note that mapping abiotic/biotic measurement fields (sex, temperature, abundance, lengths, etc.) will occur within the extendedMeasurementOrFact extension. Here this data will go from being a separate column to being condensed into the `measurementType` *and* `measurementValue` *fields.*

The obistools R package also has the map_fields function that you can use to map your dataset fields to a DwC term.

4.8 Map eMoF measurement identifiers to preferred BODC vocabulary

4.8.1 MeasurementOrFact vocabulary background

The MeasurementOrFact terms `measurementType`, `measurementValue`, and `measurementUnit` are completely unconstrained and can be populated with free text. While free text offers the advantage of capturing complex and as yet unclassified information, there is inevitable semantic heterogeneity (e.g., of spelling, wording, or language) that becomes a challenge for effective data interoperability and analysis. For example, if you were interested in finding all records related to length measurements, you would have to try to account for all the different ways "length" was recorded by data providers (length, Length, len, fork length, etc.).

> Use the OBIS Measurement Type search tool to see the diversity of `measurementTypes` that exist across published datasets in OBIS. Note that any `measurementTypeIDs` listed in this tool are **solely** for consultation purposes. In some cases codes may have been incorrectly chosen for the associated `measurementType`. You should always choose `measurementTypeIDs` based on your own data and the guidelines in this manual.

The 3 identifier terms `measurementTypeID`, `measurementValueID` and `measurementUnitID` are used to standardize the measurement types, values and units.

These three terms should be populated using controlled vocabularies referenced using Unique Resource Identifiers (URIs). For OBIS, we recommend using the internationally recognized NERC Vocabulary Server, developed by the British Oceanographic Data Centre (BODC). This server can be accessed through:

- NERC Vocabulary Server (NVS) search https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/
- Semantic Model Vocabulary Builder https://www.bodc.ac.uk/resources/vocabularies/vocabulary_builder/
- SeaDataNet facet search https://vocab.seadatanet.org/p01-facet-search

Controlled vocabularies are incredibly important to ensure data are interoperable - readable by both humans and machines and that the information is presented in an unambiguous manner. Vocabulary collections like NERC NVS2 compile vocabularies from different institutions and authorities (e.g., ISO, ICES, EUNIS), allowing you to map your data to them. In this way, you could search for a single `measurementTypeID` and obtain all related records, regardless of differences in wording or language used in the data.

Each vocabulary "term" in NVS is a concept that describes a specific idea or meaning. For consistency, we refer to individual vocabularies in NVS as **concepts**. Concepts within NVS are organized into *collections* that group concepts with commonalities (e.g. all concepts pertaining to units). Sometimes collections contain

concepts that are deprecated. Terms can be deprecated due to duplication of concepts, or when a term becomes obsolete. You should not use any deprecated concepts for any measurement ID. Deprecated concepts can be identified from lists on NVS because their identifier will have a red warning symbol, and the page for the term itself will indicate the concept is deprecated in red lettering. Unfortunately, there is currently no notification system in place to automatically warn you if a previously used concept has become deprecated. We recommend occasionally confirming that the concepts you or your institution use are still available for use.

Guidelines for populating each mesaurement ID are described below.

4.8.1.1 Guidelines to populate measurementUnitID

The `measurementUnitID` field is the easiest measurement ID field to populate. It is used to provide a URI for the unit associated with the value provided to measurementValue (e.g. cm, kg, kg/m2). This field should be populated with concepts from the **P06 collection**, BODC-approved data storage units. Documentation for this collection can be found here.

The entire vocabulary list, including deprecated terms can be found at http://vocab.nerc.ac.uk/collection/P06/current. However, we strongly recommend using https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/P06/ to avoid potentially selecting deprecated terms.

Some examples of `measurementUnits` and the associated `measurementUnitID` include:

- Metres: http://vocab.nerc.ac.uk/collection/P06/current/ULAA/
- Days: http://vocab.nerc.ac.uk/collection/P06/current/UTAA/
- Metres per second: http://vocab.nerc.ac.uk/collection/P06/current/UVAA/
- Percent: http://vocab.nerc.ac.uk/collection/P06/current/UPCT/
- Milligrams per litre: http://vocab.nerc.ac.uk/collection/P06/current/UMGL/

4.8.1.2 Guidelines to populate measurementValueID

The `measurementValueID` field is used to provide an identifying code for `measurementValues` that are **non-numerical** (e.g. sampling related, sex or life stage designation, etc.). **Note: it is not used for standardizing numeric measurements!**

Unlike `measurementUnitID`, there is more than one collection which may be used to search for and use concepts from. The collection is dependent on which type of `measurementValue` you have. See the table below for some common, non-exhaustive examples.

Type of measurementValue	Collection	Collection Documentation	Complete Vocabulary List
Sex (gender)	S10	https://github.com/nvs-vocabs/S10	<http://vocab.nerc.ac.uk/collection/S10/current/
Lifestage	S11	https://github.com/nvs-vocabs/S11	http://vocab.nerc.ac.uk/collection/S11/current/
Sampling instruments and sensors (SeaVoX Device Catalogue)	L22	https://github.com/nvs-vocabs/L22	http://vocab.nerc.ac.uk/collection/L22/current
Sampling instrument categories (SeaDataNet device categories)	L05	https://github.com/nvs-vocabs/L05	<http://vocab.nerc.ac.uk/collection/L05/current
Vessels (ICES Platform Codes)	C17	-	http://vocab.nerc.ac.uk/collection/C17/current
European Nature Information System Level 3 Habitats	C35	-	https://vocab.nerc.ac.uk/collection/C35/current/

You can also populate `measurementValueID` with references to papers or manuals that document the sampling protocol used to obtain the measurement. To do this you should use either:

- The DOI of the paper/manual
- Handle for publications on IOC's Ocean Best Practices repository, e.g. http://hdl.handle.net/11329/304

4.8.1.3 Guidelines to populate measurementTypeID

4.8.1.3.1 The P01 Collection

One of the more important collections for OBIS is the P01 collection.

Important note! **P01 codes are required for the `measurementTypeID` field**.

The P01 is a large collection with >45,000 concepts. Each concept within this collection is composed of different elements that, together, construct a label you can use for a measurement type. Frequently, concepts from the P01 collection are referred to as a "P01 code". P01 codes are used to populate the `measurementTypeID` field.

It is important to know that a semantic model, shown below, underlies each P01 code and the elements that compose them. There are 5 potential elements in this semantic model that, together, unambiguously describe a measurement type. See the P01 wheel for example of how these elements relate and combine to make one P01 code.

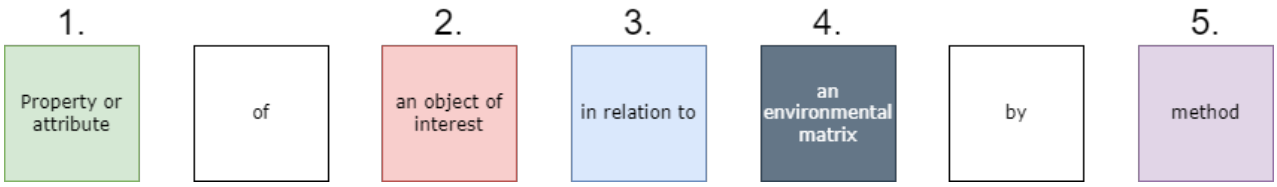


Figure 4.5: *Elements of the semantic model for P01 codes*

- **Property/attribute**: the measurement or observation of either an object of interest or a matrix, or both
- **Object of interest**: a chemical object, a biological object, a physical phenomenon, or a material object
- **In relation to**: how the measurement is related to the environment
- **Environmental matrix**: what environment the measurement is in (e.g. water body, seabed); needed for most environmental measurements, but may not be necessary for e.g. biological measurements
- **Method**: any specific methods used that are important to interpret the measurement

Note: Not every element is required, but it is important to think about each piece of the model and how it may or may not apply to your measurement. More details about this are described below in the mesaurementTypeID section.

You can use codes from other collections (e.g. P06, Q01) for `measurementValueID` and `measurementUnitID` fields, but for `measurementTypeID` you must always use a code from the P01 collection (limited exceptions, see below).

4.8.1.3.2 Selecting P01 codes for measurementTypeID When selection P01 codes, it is important to understand that each element within a P01 code is meant to describe an aspect of the measurement: what is the measurement, what is the object or entity being measured, in what environment was the measurement taken, by what kind of methods, etc.? By taking together all these elements, we are able to have a unique and specific description to differentiate one measurement from another. More documentation about the P01 code and the semantic model it is based on can be found here.

The P01 collection is found here and can be searched through the NERC vocabulary server.

You may notice when searching for measurement types related to an occurrence that specific taxonomic codes are available to you, e.g., abundance of Notommata. For OBIS, **all P01 codes should be generalized** - i.e. **do not** select species-specific codes. Instead only choose codes for "biological entities specified elsewhere" when the measurement is related to an occurrence record.

There are several ways of searching for a P01 code, but we highly recommend using the SeaDataNet P01 Facet Search. You may notice when searching for measurement types related to an occurrence that specific taxonomic options are available to you, e.g., abundance of Notommata. For OBIS, **all P01 codes should be generalized** - i.e. do not select species-specific codes. Instead, only choose codes for "biological entities specified elsewhere"! This is due to the Darwin Core Archive structure - taxonomic identification is already specified in the Occurrence table, but measurements are recorded in the ExtendedMeasurementOrFact table.

When you are comfortable and understand P01 codes, you can also use the BODC Vocabulary Builder or simply search for terms directly on the NERC Vocabulary Server.

For measurementTypes related to sampling instruments and/or methods attributes, see the Q01 collection:

- Vocabulary: http://vocab.nerc.ac.uk/collection/Q01/current/
- Search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/Q01/

Use the follow decision tree to help you select P01 codes for biological, chemical, and/or physical measurements.

4.8.2 Requesting new vocabulary terms

If you have already tried looking for a P01 code and were unable to identify a suitable code for your `measurementType` you must then request a code to be created. Before doing so, make sure you have not over filtered the search results. Then, to request a new term, your request must be submitted via:

- Submit request through the OBIS Vocabulary GitHub repository (https://github.com/nvs-vocabs/OBISVocabs/issues)
  - Requests can also be emailed to vocab.services@bodc.ac.uk if you cannot access GitHub
- Registration with the BODC Vocabulary Builder
  - Note: these requests should be based on combinations of existing concepts

We strongly recommend you use GitHub if you can, as it allows longer-term documentation, and can be relevant for other users who may be interested in the same type of code creation.

Finally, if you are unsure about whether a code fits your specific case, please feel free to ask questions to the Vocab channel on the OBIS Slack.

4.8.2.1 How to Submit a GitHub Vocabulary Request

1. Navigate to https://github.com/nvs-vocabs/OBISVocabs/issues and click on the New Issue button.
2. Click Get started
3. Fill in the title with short details of your request or issue. Then fill in the description. It is recommended to list any existing terms that are similar to your request, or concepts that are sub-components of the request.
4. *Example*: An issue was created to address difficulties in identifying P01 codes for sex rather than gender. Gender is a concept generally applied to humans, whereas "sex" is more applicable for animals. Thus the request was to either modify the current gender P01 code, or create a P01 code that specifies sex, not gender. At the time the request was issued, when users searched for a P01 term for "sex", only species-specific terms were available.

4.9 Common Data formatting issues

Contents:

4.9.1 Missing required fields

If you are a *node manager* and one of the datasets in your IPT is missing data, you should prepare a brief report to contact the data provider and outline what is missing. Get in contact with original data provider if

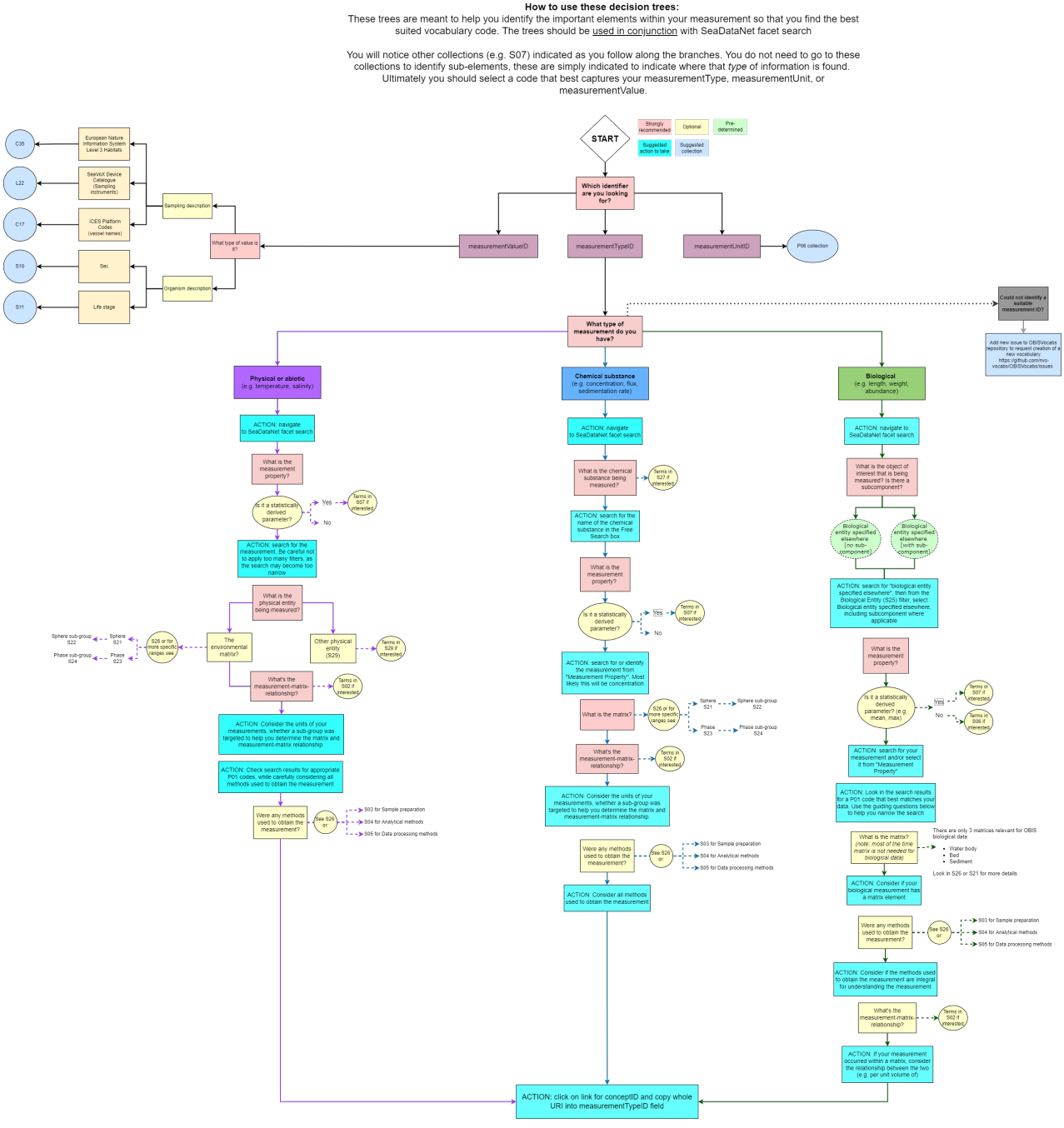


Figure 4.6: *Decision tree for measurement identifiers. Subject to updates for the measurementValue branch.*

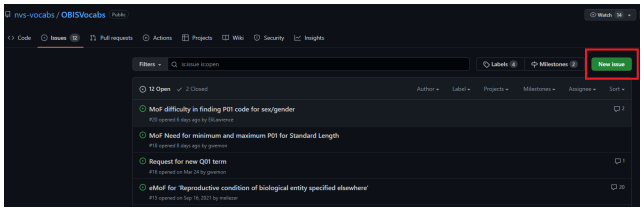


Figure 4.7: *Screenshot of how to request a new vocabulary on Github*

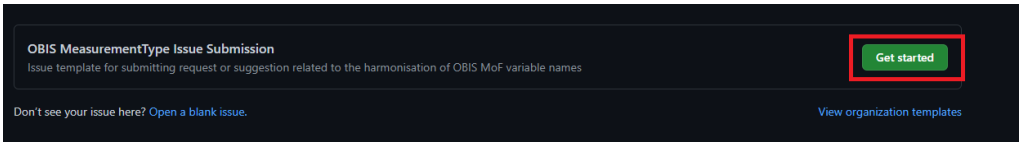


Figure 4.8: *Screenshot of submitting an issue to Github*

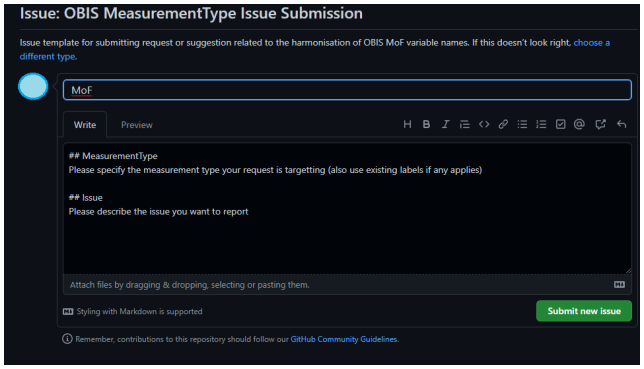


Figure 4.9: *Screenshot for how to request a new measurementType on Github*

Figure 4.10: *Example of previously requested new term on Github*

possible. If it is not possible to contact the original data provider (frequently the case for historical datasets), do your best to follow the guidelines below to fill in data. However, **do not guess or make assumptions** if you are unsure. For all fields inferred, please record notes in the `eventRemarks`, `occurrenceRemarks`, or 'identificationRemarks' field, as applicable.

If you are a *data provider* and notice or been notified that you are missing one (or more) of the eight required terms for OBIS, please proceed accordingly for each term.

To resolve missing fields marked as required by OBIS, there are several things you can do, depending on which required field is missing. Follow the guidelines below for each term.

- `occurrenceID` or `eventID`

Create a unique occurrenceID for each of your observations. These IDs can be generated by combining dates, location names, and sampling methods.

- `eventDate`

Ensure your eventDate is specified for each event, formatted according to ISO 8601 standards (e.g., YYYY-MM-DD). We have developed step by step guidelines to help you format contemporary dates and durations into ISO formatting. If your date falls outside the range of acceptable dates - i.e., historical or geological data occurring before 1583 - please follow recommendations for historical data.

For any eventDate that is inferred from literature, you should document the original date in the `verbatimEventDate` field.

- `decimalLongitude` and `decimalLatitude`

First, if you have coordinate data, make sure they are converted into decimal degrees. If you do not have specific coordinate data then you must approximate the coordinates based on locality name. You can use the Marine Regions gazetteer to search for your region of interest and obtain midpoint coordinates. Guidelines for using this tool and for dealing with uncertain geolocations can be found here. You will have to make some comments in the `georeferenceRemarks` field if you are estimating coordinates.

- `scientificName`

This field should contain only the **originally documented** scientific name down to the lowest possible taxon rank, even if there are misspellings or if it is a current synonym. Class or even Kingdom levels are accepted if more specific taxonomic levels are unknown. Comments about misspellings, etc. can be documented in the `taxonRemarks` field. Note that there may be special cases for eDNA and DNA derived data, see specific guidelines for these cases.

You may encounter challenges filling this field if the species name is based on description or if its taxonomy was uncertain at the time of sampling. For such uncertain taxonomy situations, see our guidelines here.

- `scientificNameID` (strongly recommended)

If you cannot obtain the required Life Science Identifier (LSID) from taxon matching with WoRMS then you must contact World Register Marine Species to have an LSID created for your taxon. You will need to confirm that the species is marine. OBIS does not parse LSIDs from other sources (e.g., Integrated Taxonomic Information System, Catalog of Life), but if you want to include other LSIDs alongside the WoRMS LSID, they must be specified in a predictable format.

- `occurrenceStatus`

Because occurrenceStatus is a binary field, "presence" or "absence", this field can usually be easily inferred by data. If there are associated measurements or a record of an observation, the taxon in question would be present. If a particular species/taxa is present in one sample, but missing from another, then you could identify that species as absent from the second sample.

- `basisOfRecord`

`basisOfRecord` distinguishes what type of record is in your data. For records pertaining to a collected or stored specimen, you must choose one of the following terms:

- `PreservedSpecimen`
- `FossilSpecimen`
- `LivingSpecimen`

For records pertaining to an observation in the wild, you should use:

- `HumanObservation` (e.g., observation in the wild)
- `MachineObservation` (e.g., photograph, acoustic detection, etc.)
- `MaterialSample` (e.g. DNA sequences, DNA detection)

For records pertaining to literature data, `basisOfRecord` should always reflect the evidence upon which the Occurrence record was based. For example, a researcher's record based on photographs should specify `MachineObservation`, otherwise specifications should be `HumanObservation` (see relevant GitHub discussion).

For specifics on when to use each of these and which other fields should be populated along with them, see the guidelines on record-level terms.

4.9.2 Temporal: Dates and times

The date and time at which an event took place or an occurrence was recorded goes in `eventDate`. This field uses the ISO 8601 standard. OBIS recommends using the extended ISO 8601 format with hyphens. Note that all dates in OBIS become translated to UTC during the quality control process implemented by OBIS. Formatting your dates correctly ensures there will be no errors during this process.

ISO 8601 dates can represent moments in time at different resolutions, as well as time intervals, which use / as a separator. Date and times are separated by `T`. Timezones can be indicated at the end by using + or - the number of hours offset from UTC. If no timezone is indicated, then the time is assumed to be local time. When a date/time is recorded in UTC, a Z should be added at the end. Times must be written in the 24-hour clock system. If you do not know the time, you do not have to provide it. Please do not indicate unknown times as "00:00" as this indicates midnight.

Not every piece of time information is necessary, but a generalization of how to format dates and times looks like:

```
YYYY-MM-DDT[hh]:[mm]:[ss] [+/-XX OR Z]
```

Some specific examples of acceptable ISO 8601 dates are:

Dates:

- 1948-09-13
- 1993-01/02
- 1993-01
- 1993

Dates with Specific Times:

- 1973-02-28T15:25:00
- 2008-04-25T09:53

Dates with Time Zones:

- 2005-08-31T12:11+12
- 2013-02-16T04:28Z

Date and Time Intervals:

- 1993-01-26T04:39+12/1993-01-26T05:48+12

It is important to note that although ISO 8601 also supports ordinal dates (YYYY-DDD) and week dates (YYYY-Www-D), these formats are not supported by OBIS. Additionally, ISO 8601 guidelines for durations should not be used. Durations for an event (e.g., length of observation) can instead be indicated with the DwC terms startDayOfYear and endDayOfYear. Durations refer to the actual length of time an event (e.g., occurrence) occurred, whereas intervals indicate the time period during which an event was recorded.

A note about intervals... Take care when entering date intervals as, for example, entering 1960/1975-08-04 indicates that the event or observation started any time in 1960, and ended any time on 1975-08-04. If you know the exact date and time, you should specify that information. This also helps for continuous samplings and time-series integrated datasets.

If you have a mix of dates and times for different aspects of a sampling event, you can embed this information in the Event Core table using hierarchies of date structure. To do this, you can use separate records for events, and specify each event date individually. See example.

For uncertainty regarding the date of the event, see guidelines.

4.9.2.1 Tips

To ensure your date is formatted correctly, it may be easiest to begin by populating the `year`, `month`, and `day` fields first. If the specific time of sampling is known, populate that into `eventTime` as well. When you fill these fields, we recommend ensuring the numbers are encoded as Text, not as General or numeric as Excel often tries to interpret what it thinks the content "should" be. Otherwise you may run into problems with Excel auto formatting your numbers in ways you don't want. You can do this by highlighting the cells of interest, navigating to the Number Format on the Home ribbon and selecting "Text". Be careful when you do this change of format, as some columns (e.g. time) may become formatted into a decimal or other unexpected format.

Then you can use Excel to concatenate each field together, adding the time zone at the end, using the general format:

```
=CONCAT(YEAR, "-", MONTH, "-", DAY, "T", EVENTTIME, TIMEZONE)
```

> Note You can also use the Canadensys date parsing tool to help you convert dates or parse them into component parts.

A caution about dates and Excel: Excel is unfortunately notorious for causing issues in saving dates. The Data Carpentries have produced this exercise which demonstrates how Excel interprets dates and numbers, sometimes converting numbers into dates and vice versa. This exercise is simply a demonstration of Excel - it does not provide advice on formatting dates for OBIS.

Date formats in Excel can be very dependent on your computer system region custom and not all of them have the ISO 8601 format included. Therefore you can type the date in the requested format but it will automatically revert the format according to your Windows system region settings. You can change your system region by: navigating to Control Panel > All Control Panel Items > Region and then select "English (United States)" or "English (United Kingdom)". The YYYY-MM-DD format will appear among the choices within the Format cells - Date options.

If your computer language is not set to English, you may encounter additional issues with Excel. It may change the format of your date even after you save the document. Changing your computer system's language to English can help, but you may still run into issues. You may also try using other office management softwares, like LibreOffice which in this case is more friendly. In general, we advise you to be very careful when formatting the `eventDate` field, and to select the "Text" formatting (as above) and to save your file as a .CSV.

It is good practice to place the verbatim event date/time description into the `verbatimEventDate` field. Any modifications you make during data formatting should be recorded in the `eventRemarks` field, and we recommend taking good notes in a personal reference file.

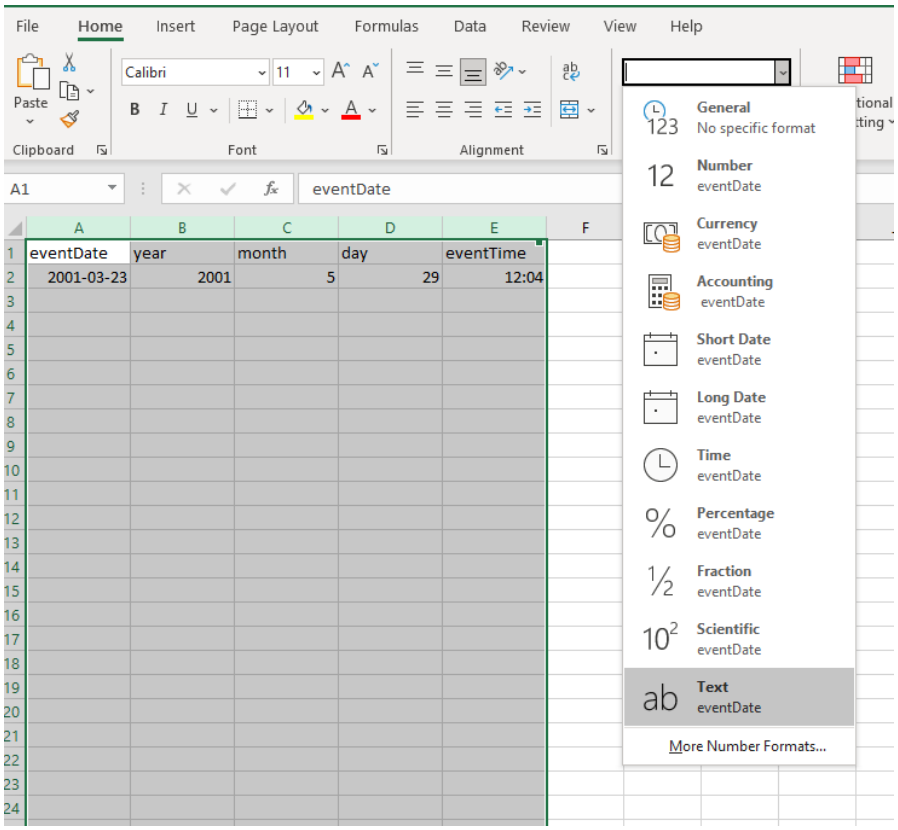


Figure 4.11: *Screenshot of how to change data type in Excel*

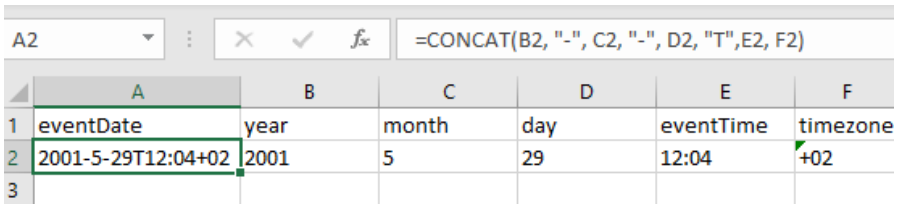


Figure 4.12: *Example of how to concatenate dates in Excel*

This video provides a demonstration of how you can format dates to ISO 8601, including how to resolve difficulties you might run into (some which are also discussed below).

4.9.2.2 How to handle mixed date information

When the sampling date is unclear due to a mix of date types and durations, you can use the hierarchical event structure in the Event core to help format and associate dates (and other information) with the correct sampling event. Often times it is useful to break up each type of event into separate records according to the event hierarchy.

Let's look at a fictional dataset to demonstrate this. In this example, there is a project called Maple. This project has a cruise that takes place in May and June which takes samples at three different sites. The project has data for three years, beginning in 1993. How do we capture all this information in our dataset? And how do we format `eventDate` to reflect these different times and durations?

We can embed this information using a different `eventID` and `parentEventID` for each level within the project - the reoccurring cruise, the station sites, and then the samples themselves. We have added a column here for "Event Type" for the sake of example. At this time, there is no Darwin Core event type term, however there is discussion for its creation.

Our example dataset would then look like the following:

parentEventID	eventID	Event type	eventDate	eventRemarks
	MAPLE_1993_crs	cruise		This is the first year
	MAPLE_1994_crs	cruise		This is the second year
	MAPLE_1995_crs	cruise		This is the third year
MAPLE_1993_crs	MAPLE_1993_crs_st1	station		This is the first year, first site
MAPLE_1993_crs	MAPLE_1993_crs_st2	station		This is the first year, second site where samples were taken over two days
MAPLE_1993_crs	MAPLE_1993_crs_st3	station		This is the first year, third site
MAPLE_1993_crs_st1	MAPLE_1993_crs_st1_s1	sample	1993-05-05T10:13-04	This is the first year, first site, first sample
MAPLE_1993_crs_st1	MAPLE_1993_crs_st1_s2	sample	1993-05-05T10:38-04	This is the first year, first site, second sample
MAPLE_1993_crs_st2	MAPLE_1993_crs_st2_s1	sample	1993-05-19T23:40-04	This is the first year, second site, first sample
MAPLE_1993_crs_st2	MAPLE_1993_crs_st2_s2	sample	1993-05-20T01:24-04	This is the first year, second site, second sample
MAPLE_1993_crs_st3	MAPLE_1993_crs_st3_s1	sample	1993-06-01T09:21-04	This is the first year, third site, first sample
MAPLE_1993_crs_st3	MAPLE_1993_crs_st3_s2	sample	1993-06-01T09:57-04	This is the first year, third site, second sample

You can see that the eventDate for the parent events does not need to be provided - only the dates for the actual samples are required.

4.9.3 Historical data

OBIS recognizes the difficulties in formatting historical, archaeological and paleontological data series. Time-series are often misinterpreted, difficult to formulate, or confined to humanities disciplines rather than the "science" community. This kind of data is sometimes seen as "specialist" or "niche" when it comes to sharing in globally accepted databases that are accessible and recognised in academic, research and scientific forums.

Some of the nuances associated with historical data have to do with the change in calendar systems, from the Julian calendar to the currently used (by most countries) Gregorian calendar metric system. This change was implemented in 1582, so any datasets with data representing periods that predate this year **must** be checked and converted to the standard Gregorian calendar system. Additionally, there is no year zero, only -1 and 1, where -1 is BCE (Before Common Era) and 1 is CE (Common Era). This can make interpretation of historical dates more challenging as such dates will need to be converted to align with ISO 8601 standards.

To accommodate such challenges the OBIS Historical Data Project Team recommends the following:

- Always populate `verbatimEventDate` with the originally documented date so that it can be preserved. Place converted dates that align to ISO 8601 in the `eventDate` field, and document the changes you made to the original in `eventRemarks`.

- When the exact date is unknown, provide a date range, e.g. the period 21 November 1521 to 29 August 1612 records as 1521-11-21/1612-08-29.
- For archaeological data, use terms from the Darwin Core class GeologicalContext and/or the Chronometric Age Extension. GeologicalContext terms can be used to capture information such as periods or ages, however the Chronometric Age extension allows for more thorough descriptions of the time period and can link to the Event core table. For such records, `eventDate` would be populated with the date of collection.
- If the historical record contains uncertain or sensitive location information, generalize the location information using polygons or lines as described in this Manual.

For historical data originating from old records, such as ship logs or other archival records, we understand there can be a variety of issues in interpreting and formatting data according to DwC standards. If you need further help with historical data formatting, we recommend submitting a Github issue, or contacting the OBIS-OPI node who focuses on Oceans Past historical, archaeological, and paleontological data series.

4.9.4 Spatial

4.9.4.1 Converting Coordinates

All coordinates provided in the `decimalLatitude` or `decimalLongitude` fields in OBIS must be in decimal degrees. To convert coordinates from degrees-minutes-seconds into decimal degrees, you can use this Coordinate Conversion tool that OBIS has developed. This tool will convert any coordinate (or list of coordinates on a separate line) in a degrees-minutes-seconds format into decimal degrees, even partial coordinates. To use it, simply copy and paste your coordinates into the box provided and click Convert. For example:

Watch this video for a demonstration on use of this tool.

If your coordinates are in UTMs, then coordinate conversion can be a bit trickier. We suggest using the following conversion tool to convert from UTM to decimal degrees. Note it is very important to ensure you have the correct UTM zone, otherwise the coordinate conversion will be incorrect. You can use this ArcGIS map tool to visually confirm UTM zones.

4.9.4.2 Geographical format conversion

In OBIS, the spatial reference system to be documented in `geodeticDatum` is EPSG:4326 (WGS84). If your spatial data are not already in this format, you may have to convert it. To do this there are a few approaches: QGIS (or ArcGIS), R, or Python. We provide some short guidance for each, however if you are struggling to convert your data to WGS84 please contact helpdesk@obis.org or send a message on the OBIS Slack.

4.9.4.2.1 QGIS You can load a .csv file containing your coordinates to be reprojected into QGIS. Opening a new project, first set the global projection to WGS84 EPSG:4326. In the bottom right corner, click the Project Properties to change the Project Coordinate Reference System (CRS). A pop up window will allow you to search for and select WGS84 EPSG:4326. Click OK.

To load your .csv file containing the longitude and latitude coordinates, go to Layer < Add Layer < Add Delimited Text layer...

A popup window will allow you to browse and select your .csv file. Open the `Geometry Definition` portion of the window and map the field containing longitude values to the `X field` and latitude to the `Y field`. Select the CRS that these coordinates were recorded as from the drop down menu. Then click `Add` and close the window.

Go to Vector < Geometry Tools < Add Geometry Attributes

Make sure the input layer is your coordinate file. Under the `Calculate using`, select Project CRS (because we set the Project CRS to the desired projection). Click `Run`. This will create a new layer with an additional two columns called Xcoord (longitude) and Ycoord (latitude). These fields contain the coordinates in the desired

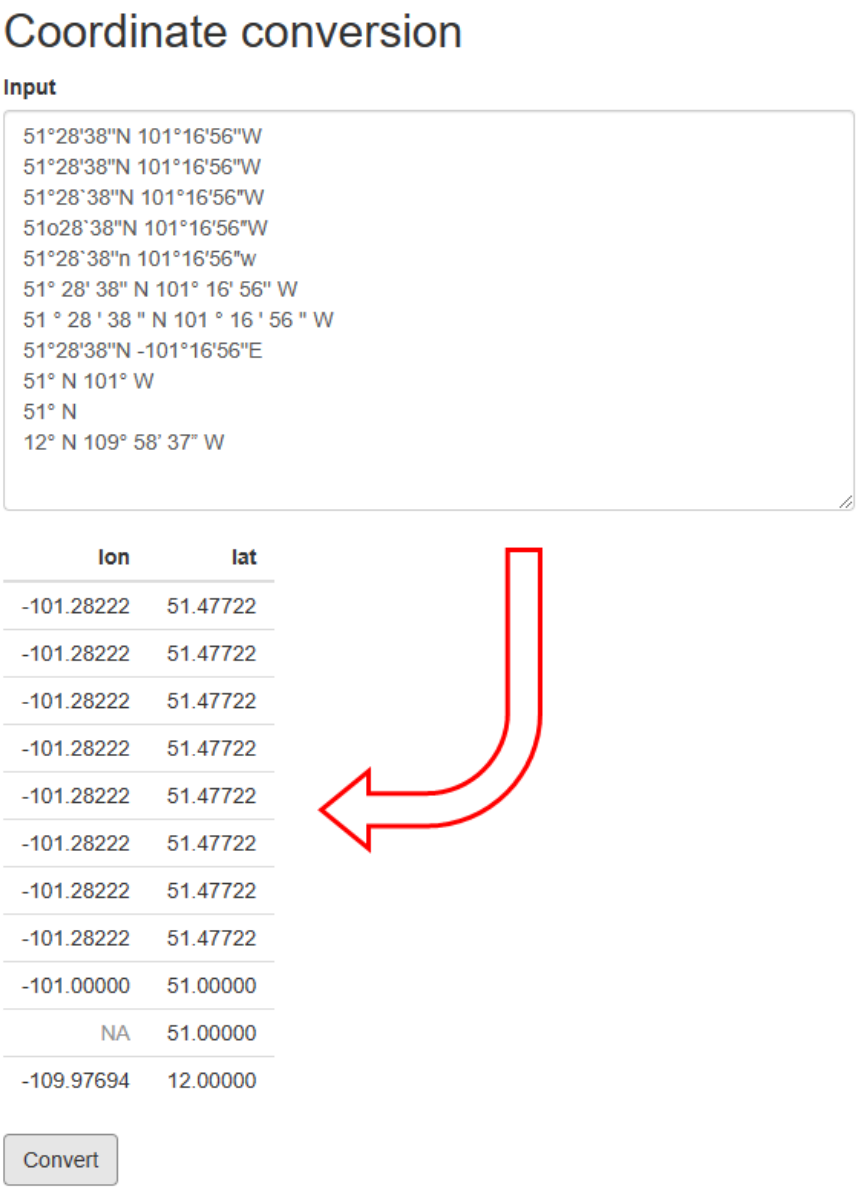


Figure 4.13: *Screenshot of how to use the OBIS coordinate converter*

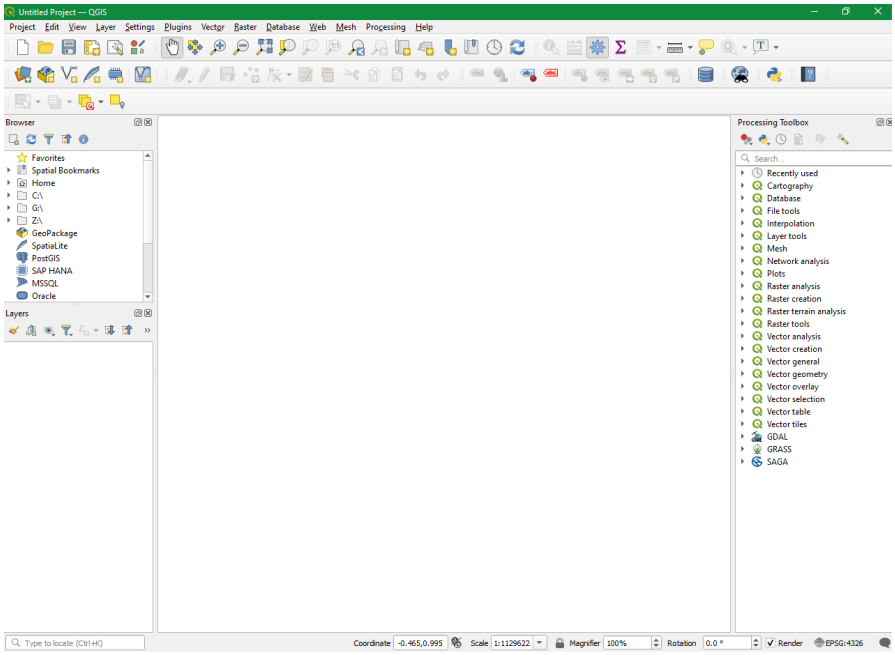


Figure 4.14: *Screenshot of QGIS interface*

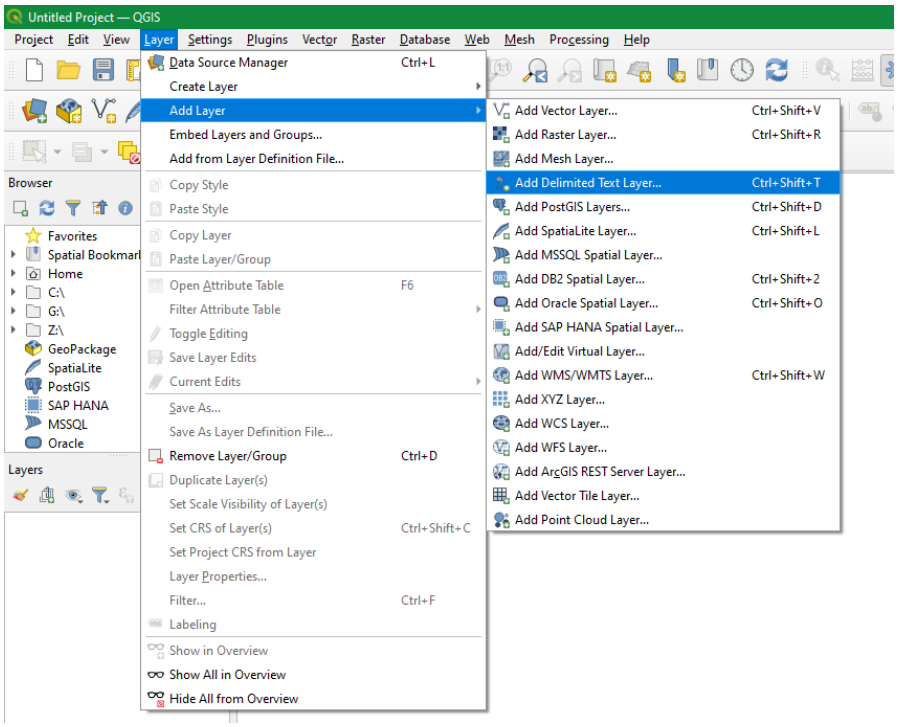


Figure 4.15: *How to add a .csv with coordinate data in QGIS*

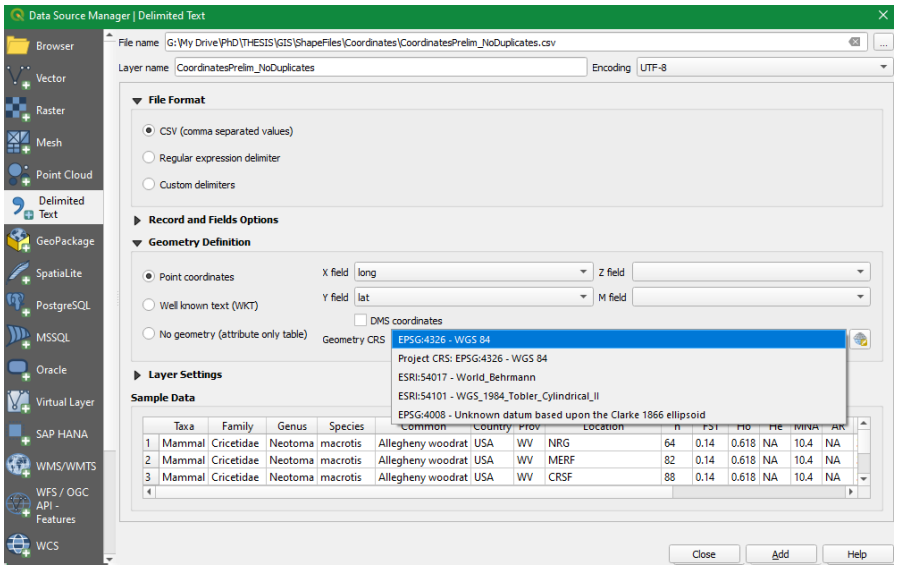


Figure 4.16: *Screenshot showing how to specify CRS of a .csv file when importing into QGIS*

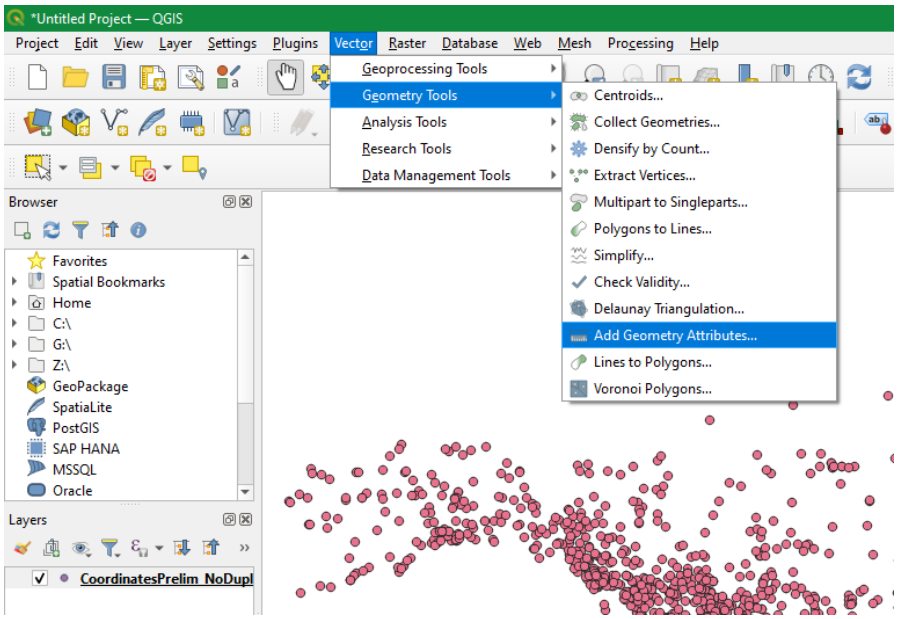


Figure 4.17: *Screenshot showing where to find the Geometry Attributes in QGIS menu*

projection (i.e., WGS84). You can view these columns by right clicking and opening the layer's attribute table. To export the file, right click the layer and click Make Permanent. Then save the .csv.

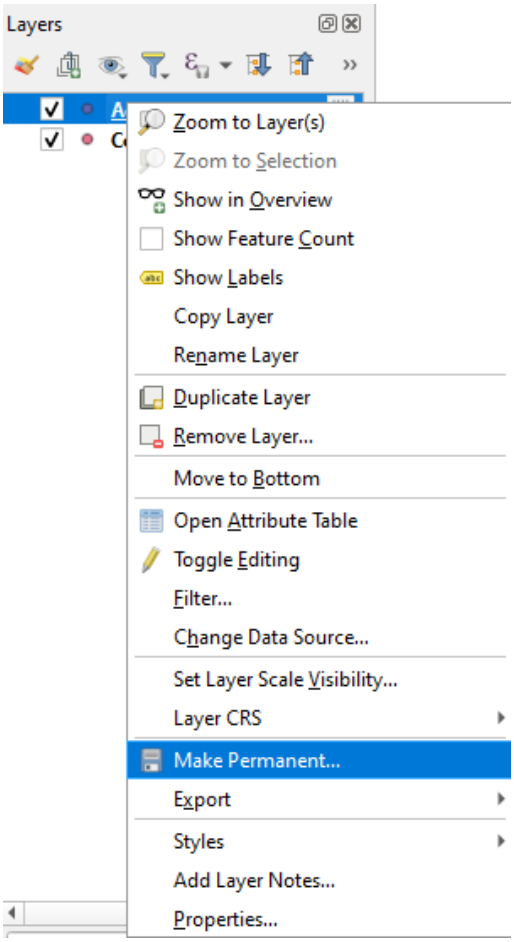


Figure 4.18: *Screenshot showing how to save a temporary layer in QGIS for export*

For more details see this QGIS guide on reprojection.

4.9.4.2.2 R To reproject coordinates in R, you can use functions in the `sf` package. A thorough tutorial using this package can be found here.

4.9.4.2.3 Python You also have the option to reproject data using the Python library Geopandas. In this package there is a utility called `to_crs` that will reproject data. A tutorial to do this can be found here.

4.10 Examples: ENV-DATA and DNA derived data

Contents

- Marine birds abundance & distribution
- Marine mammals abundance & distribution
- Marine turtles abundance & distribution
  - Survey & sighting data
- Microbes biomass & diversity
- Phytoplankton biomass & diversity
- Seagrass cover & composition
- Zooplankton biomass & diversity

Special data types:

- eDNA & DNA derived data
  - eDNA data from Monterey Bay, California
  - 16S rRNA gene metabarcoding data of Pico- to Mesoplankton
- Acoustic, imaging, or other multimedia data
- Tracking data
- Habitat

4.10.0.1 Fish abundance and distribution

(example coming soon)

4.10.0.2 Hard coral cover and composition

(example coming soon)

4.10.0.3 Invertebrates abundance and distribution

(example coming soon)

4.10.0.4 Macroalgae canopy cover and composition

In this section we will encode a fictional macroalgal survey dataset into Darwin Core using the ENV-DATA approach, i.e. using an Event core with an Occurrence extension and an extendedMeasurementOrFact extension.

Figure: A fictional macroalgae survey with a single site, multiple zones, quadrats, and different types of transects.

Event core:

First we can create the Event core table by extracting all events in a broad sense and populating attributes such as time, location, and depth at the appropriate level. The events at the different levels are linked together using `eventID` and `parentEventID`. As the survey sites has a fixed location we can populate `decimalLongitude` and `decimalLatitude` at the top level event. The zones have different depths, so `minimumDepthInMeters` and `maximumDepthInMeters` are populated at the zone level. Finally, as not all sampling was done on the same day, `eventDate` is populated at the quadrat and transect level.

eventID	parentEventID	eventDate	decimalLongitude	decimalLatitude	minimumDepthInMeters	maximumDepthInMeters
site_1			54.7943	16.9425		
zone_1	site_1				0	0
zone_2	site_1				0	5
zone_3	site_1				5	10
quadrat_1	zone_1	2019-01-02				
transect_1	zone_2	2019-01-03				
transect_2	zone_3	2019-01-04				

Occurrence extension:

Next we can construct the Occurrence extension table. This table has the scientific names and links to the World Register of Marine Species in `scientificNameID`. The first column of the table references the events in the core table (see `quadrat_1` for example highlighted in green).

id	occurrenceID	scientificName	scientificNameID
quadrat_1	occ_1	Ulva rigida	urn:lsid:marinespecies.org:taxname:145990
quadrat_1	occ_2	Ulva lactuca	urn:lsid:marinespecies.org:taxname:145984
transect_1	occ_3	Plantae	urn:lsid:marinespecies.org:taxname:3
transect_1	occ_4	Plantae	urn:lsid:marinespecies.org:taxname:3
transect_2	occ_5	Gracilaria	urn:lsid:marinespecies.org:taxname:144188
transect_2	occ_6	Laurencia	urn:lsid:marinespecies.org:taxname:143914

extendedMeasurementOrFact (eMoF) extension:

And finally there is the MeasurementOrFact extension table, which has attributes of the zones (shore height), the quadrats (surface area), the transects (surface area and length), and the occurrences (percentage cover and functional group). Attributes of occurrences point to the Occurrence extension table using the `occurrenceID` column (see `occ_1` and `occ_2` highlighted in blue and orange). Note that besides NERC vocabulary terms we are also referencing the CATAMI vocabulary for macroalgal functional groups.

id	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID
zone_1		shore height	?	high	?		
quadrat_1		surface area	P01/current/AREABOTS	0.25		m2	P06/current/UMSQ
quadrat_1	occ_1	cover	P01/current/SDBIOL24	40		percent	P06/current/UPCT
quadrat_1	occ_2	cover	P01/current/SDBIOL24	60		percent	P06/current/UPCT
transect_1		surface area	P01/current/AREABOTS	60		m2	P06/current/UMSQ
transect_1		length	P01/current/LENTRANSECT	30		m	P06/current/ULAA
transect_1	occ_3	functional group	?	sheet-like red	CATAMI:80300925		
transect_1	occ_4	functional group	?	filamentous brown	CATAMI:80300931		
transect_1	occ_3	cover	P01/current/SDBIOL24	10		percent	P06/current/UPCT
transect_1	occ_4	cover	P01/current/SDBIOL24	40		percent	P06/current/UPCT
transect_2	occ_5	cover	P01/current/SDBIOL24	10		percent	P06/current/UPCT
transect_2	occ_6	cover	P01/current/SDBIOL24	60		percent	P06/current/UPCT

4.10.0.5 Mangroves cover and composition

(example coming soon)

4.10.0.6 Marine birds abundance and distribution

The example for ENV-DATA collected with marine bird sightings/occurrences is based on the dataset "RV Investigator Voyage IN2017_V02 Seabird Observations, Australia (2017)". In this dataset, seabird sightings were recorded continuously during daylight hours during a voyage to recover and redeploy moorings at the SOTS site, southwest of Tasmania, Australia, in March 2017. Observations were made from c.30 minutes before sunrise to c.30 minutes after sunset, extending to 300m in the forward quadrant with the best viewing conditions. There were 1200 observations from 38 species of birds along with 3 cetacean species and one seal. This example will focus on the ENV-DATA associated with the bird sightings. The most frequently sighted bird species were *Puffinus tenuirostris* (Short-tailed Shearwater) and *Pachyptila turtur* (Fairy Prion).

For this dataset, human observation recorded individual bird sightings (thus, each specimen is a single occurrence). The dataset contains abiotic measurements (ENV-DATA) which are related to each individual sighting, instead of an entire sample. Therefore, we can create an Occurrence core with an eMoF extension that contain the abiotic environmental measurements or facts.

Occurrence core:

The Occurrence core is populated with the occurrence records of seabirds sighted during the RV voyages. Occurrence details and scientific names are provided here. All birds were observed above sea level, all `minimumDepthInMeters` and `maximumDepthInMeters` values equal zero.

occurrenceID	eventDate	institutionCode	collectionCode	basisOfRecord	recordedBy	organismQuantity	organismQuantityType	sex	decimalLatitude	decimalLongitude	geodeticDatum	coordinateUncertaintyInMeters	footprintWKT
in2017_v02_20170317_v02_00998-1	2017-03-17 01:07:00	Australasian Seabird Group, BirdLife Australia	in2017_v02_Human Observation	HumanObservation	EBWiG+GRC+TAH	individuals	present		-43.40741	147.45576	200	0.0018	POINT (147.45576 -43.40741)
in2017_v02_20170319_v02_00998-2	2017-03-19 22:26:00	Australasian Seabird Group, BirdLife Australia	in2017_v02_Human Observation	HumanObservation	EBWiG+GRC+TAH	individuals	present		-45.98644	142.1445	200	0.0018	POINT (142.14450 -45.98644)
in2017_v02_20170317_v02_00998-3	2017-03-17 02:38:00	Australasian Seabird Group, BirdLife Australia	in2017_v02_Human Observation	HumanObservation	EBWiG+GRC+TAH	individuals	present		-43.40728	147.45549	200	0.0018	POINT (147.45549 -43.40728)

extendedMeasurementOrFact (eMoF) extension:

As shown in previous examples, the MeasurementOrFact extension table contains abiotic measurements or facts corresponding to an occurrence / sighting. Individual sightings and abiotic measurements are linked with `occurrenceID`. In the example dataset, the ENV-DATA consist of measurements taken during the moorings deployment at the SOTS site, at the time of the marine bird sightings. In addition to NERC vocabulary terms, authors also referenced the Australian Ocean Data Network (AODN) Discovery Parameter Vocabulary for *Sea-floor depth (m)* and *Sea Surface Temperature* as `measurementType`. NERC equivalents to the AODN terms are added as additional MeasurementOrFact (MoF) records.

occurrenceID	measurementID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID
in2017_v02_20170317_v02_00998-1	in2017_v02_20170317_v02_00998-depth	Sea-floor depth (m)	http://vocab.aodn.org.au/def/discovery_parameter/entity/574	73.0313	NA	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA
in2017_v02_20170317_v02_00998-2	in2017_v02_20170317_v02_00998-depth	Sea-floor depth	http://vocab.nerc.ac.uk/collection/P01/current/MBANZZZZ/	73.0313	NA	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA
in2017_v02_20170317_v02_00998-3	in2017_v02_20170317_v02_00998-air__pressure	Air Pressure (hPa)	http://vocab.nerc.ac.uk/collection/P01/current/CAPHZZ01	1024.91385	NA	hPa	http://vocab.nerc.ac.uk/collection/P06/current/HPAX
in2017_v02_20170317_v02_00998-1	in2017_v02_20170317_v02_00998-air__temp	Atmospheric temperature (deg C)	http://vocab.nerc.ac.uk/collection/P01/current/CTMPZZ01	15.3	NA	degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA
in2017_v02_20170317_v02_00998-1	in2017_v02_20170317_v02_00998-wov__sea__state	Sea state	http://vocab.nerc.ac.uk/collection/C39/current/	moderate 1.25 - 2.5 m	http://vocab.nerc.ac.uk/collection/C39/current/4/		
in2017_v02_20170317_v02_00998-1	in2017_v02_20170317_v02_00998-sea__surface__temp	Sea surface temperature	http://vocab.aodn.org.au/def/discovery_parameter/entity/97	17.32	NA	degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA
in2017_v02_20170317_v02_00998-2	in2017_v02_20170317_v02_00998-sea__surface__temp	Sea surface temperature	http://vocab.nerc.ac.uk/standard__name/sea__surface__temperature/	17.32	NA	degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA
in2017_v02_20170317_v02_00998-3	in2017_v02_20170317_v02_00998-wind__direction	Wind direction (deg)	http://vocab.nerc.ac.uk/collection/P01/current/EWDAZZ01	283	NA	degrees	http://vocab.nerc.ac.uk/collection/P06/current/UABB
in2017_v02_20170317_v02_00998-1	in2017_v02_20170317_v02_00998-wind__speed	Wind Speed (knt)	http://vocab.nerc.ac.uk/collection/P01/current/ESSAZZ01	5.49	NA	Knots (nautical miles per hour)	http://vocab.nerc.ac.uk/collection/P06/current/UKNT

4.10.0.7 Marine mammals abundance and distribution

In this section we will explore how to encode a survey data set into Darwin Core using the ENV-DATA approach. As an example, sections of the actual data set of CETUS: Cetacean monitoring surveys in the Eastern North Atlantic, is used.

Figure: A representation of the observation events of CETUS: Cetacean monitoring surveys in the Eastern North Atlantic, presenting the route **Madeira** as a site with three cruises (zones). Each **Cruise** is divided into different **Transects** and each transect contains a number of **Positions**.

Event core:

Create the Event core table by extracting all events and populating attributes. As in the previous example, the events at the different levels are linked together using `eventID` and `parentEventID`. As the survey observations

were made at locations of cetacean sightings instead of fixed locations, we can populate `footprintWKT` and `footprintSRS` as location information. Not all sampling was done on the same day, therefore `eventDate` is populated at the transect level.

eventID	parentEventID	eventDate	footprintWKT	footprintSRS
Madeira		2012-07/2017-09	POLYGON ((-16.74 31.49, -16.74 41.23, -8.70 41.23, -8.70 31.49, -16.74 31.49))	EPSG:4326
Madeira:Cruise-001	Madeira	2012-07	MULTIPOINT ((-8.7 41.19), (-9.15 38.7))	EPSG:4326
Madeira:Cruise-002	Madeira	2012-07	MULTIPOINT ((-9.15 38.7), (-16.73 32.74))	EPSG:4326
Madeira:Cruise-003	Madeira	2012-07	MULTIPOINT ((-16.73 32.74), (-9.15 38.7))	EPSG:4326

Occurrence extension:

Construct the Occurrence extension table with the scientific names and links to the World Register of Marine Species in `scientificNameID`. The first column of the table references the events in the core table (see `Madeira:Cruise-001` highlighted in green).The `occurrenceID` corresponds to the Position of the observation (see `Transect-01:Pos-0001` and `CIIMAR-CETUS-0001` highlighted in blue, or `Transect-01:Pos-0002` and `CIIMAR-CETUS-0002` highlighted in orange).

id	occurrenceID	scientificNameID	scientificName
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	urn:lsid:marinespecies.org:taxname:2688	Cetacea
Madeira:Cruise-001:Transect-01:Pos-0002	CIIMAR-CETUS-0002	urn:lsid:marinespecies.org:taxname:2688	Cetacea
Madeira:Cruise-001:Transect-01:Pos-0003	CIIMAR-CETUS-0003	urn:lsid:marinespecies.org:taxname:2688	Cetacea
Madeira:Cruise-001:Transect-02:Pos-0004	CIIMAR-CETUS-0004	urn:lsid:marinespecies.org:taxname:2688	Cetacea
Madeira:Cruise-001:Transect-02:Pos-0005	CIIMAR-CETUS-0005	urn:lsid:marinespecies.org:taxname:2688	Cetacea
Madeira:Cruise-001:Transect-02:Pos-0006	CIIMAR-CETUS-0006	urn:lsid:marinespecies.org:taxname:2688	Cetacea
Madeira:Cruise-001:Transect-02:Pos-0007	CIIMAR-CETUS-0007	urn:lsid:marinespecies.org:taxname:2688	Cetacea

extendedMeasurementOrFact (eMoF) extension:

And finally, the extendedMeasurementOrFact extension table has attributes of the zones (such as Vessel speed and Vessel Heading), the Transects (such as Wave height and Wind speed), and the Positions (such as Visibility and the Number of smaal/big ships >20m). Attributes of Positions point to the Occurrence extension table using the `occurrenceID` column (see `Transect-01:Pos-0001` and `Transect-01:Pos-0002` highlighted in blue and orange, respectively).

id	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementUnit	measurementUnitID
Madeira:Cruise-001		Vessel name	Q01/current/Q0100001	Monte da Guia		
Madeira:Cruise-001:Transect-01		Length of the track	P01/current/DSRNCV01	39.75	km	P06/current/ULKM
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	Visibility		2000-4000	Meters	P06/current/ULAA
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	Wind speed	P01/current/WMOCWFBF	1	Beaufort scale	
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	Wave height		2	Douglas scale	
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	Number of big ships (>20m)		3		
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	Vessel heading	P01/current/HDNGGP01	206	Degrees	P06/current/UAAA
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	Number of small ships (<20m)		0		
Madeira:Cruise-001:Transect-01:Pos-0001	CIIMAR-CETUS-0001	Vessel speed	P01/current/APSAGP01	16	Knots (nautical miles per hour)	P06/current/UKNT

4.10.0.8 Marine turtles abundance and distribution

4.10.0.8.1 Survey and sighting data This section deals with encoding survey and/ or sighting data of sea turtles into Darwin Core using the ENV-DATA approach. Extracts from the actual data set of

Presence of sea turtles collected through Fixed-Line-Transect monitoring across the Western Mediterranean Sea (Civitavecchia-Barcelona route) between 2013 and 2017, are used as an example.

Event core:

The Event core is created by extracting all sighting events and populating the attributes at each event. The events at the different levels are linked together using `eventID` and `parentEventID`. In the example dataset, turtle sightings have been recorded since 2007, along a ferry route between Italy and Spain, as part of the monitoring project FLT Med Net (Fixed Line Transect Mediterranean monitoring Network). Turtle sighting locations can be given by populating the fields `footprintWKT` and `footprintSRS` with location information. Sightings were recorded at different dates, therefore `eventDate` is populated at the transect level.

id	modified	datasetID	datasetName	eventID	parentEventID	eventID	eventDate	eventRemarks	minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	geodeticDatum	coordinateUncertaintyInMeters	footprintWKT	footprintSRS
TURTLE_2013_CBAR_0043	2021-05-05 07:59:08	https://marineinfo.org/id/dataset/6403	Presence of sea turtles collected through Fixed-Line-Transect monitoring across the Western Mediterranean Sea	TURTLE_CBAR_2013		2013-04-03T05:30:00+02:00/2013-04-03T16:00:00+02:00	transect	0	0	41.26179	9.73265	EPSG:4326	2226	70.2874	LINESTRING (7.602633333333 41.24378333333, 2.263897 41.279816)	EPSG:4326
TURTLE_2013_CBAR_0045	2021-05-05 07:59:08	https://marineinfo.org/id/dataset/6403	Presence of sea turtles collected through Fixed-Line-Transect monitoring across the Western Mediterranean Sea	TURTLE_CBAR_2013		2013-04-18T05:22:00+02:00/2013-04-18T15:53:00+02:00	transect	0	0	41.30371	8.73657	EPSG:4326	2226	20.0359	LINESTRING (7.636983333333 41.32418333333, 2.236159 41.283244)	EPSG:4326
TURTLE_2013_CBAR_0045:0001	2021-05-05 07:59:08	https://marineinfo.org/id/dataset/6403	Presence of sea turtles collected through Fixed-Line-Transect monitoring across the Western Mediterranean Sea	TURTLE_CBAR_2013_0045		2013-04-18T05:55:00+02:00	sample	0	0	41.3228	7.4984	EPSG:4326			POINT	EPSG:4326
TURTLE_2013_CBAR_0045:0002	2021-05-05 07:59:08	https://marineinfo.org/id/dataset/6403	Presence of sea turtles collected through Fixed-Line-Transect monitoring across the Western Mediterranean Sea	TURTLE_CBAR_2013_0045		2013-04-18T08:35:00+02:00	sample	0	0	41.322845	5.995345	EPSG:4326			POINT	EPSG:4326

Occurrence extension:

The Occurrence extension contain details regarding the sighted animals and include `scientificName` and the links to the World Register of Marine Species in `scientificNameID`. The `EventID` references the events as in the Event core. This table further provides information on the `basisOfRecord` and `occurrenceStatus`.

EventID	occurrenceID	datasetID	collectionCode	basisOfRecord	catalogNumber	recordedBy	occurrenceStatus	scientificNameID	scientificName	kingdom	scientificNameAuthorship
TURTLE_CBAR_0001	TURTLE_CBAR_0001	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_0001	HumanObservation	Adult_TURTLE_17	Barbara Campana | Miriam Paraboschi | Erica Ercoli | Erica Arcangeli | Cristina Berardi | Lucilla Giulietti | Claudia Boccardi	absent	urn:lsid:marinespecies.org:taxname:136999	Cheloniidae	Animalia	Oppel, 1811
TURTLE_CBAR_0004	TURTLE_CBAR_0004	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_0004	HumanObservation	Adult_TURTLE_17	Barbara Campana | Miriam Paraboschi | Erica Ercoli | Erica Arcangeli | Cristina Berardi | Lucilla Giulietti | Claudia Boccardi	absent	urn:lsid:marinespecies.org:taxname:136999	Cheloniidae	Animalia	Oppel, 1811
TURTLE_CBAR_0005	TURTLE_CBAR_0005	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_0005	HumanObservation	Adult_TURTLE_17	Barbara Campana | Miriam Paraboschi | Erica Ercoli | Erica Arcangeli | Cristina Berardi | Lucilla Giulietti | Claudia Boccardi	present	urn:lsid:marinespecies.org:taxname:137205	Caretta caretta	Animalia	Linnaeus, 1758
TURTLE_CBAR_0006	TURTLE_CBAR_0006	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_0006	HumanObservation	Adult_TURTLE_17	Barbara Campana | Miriam Paraboschi | Erica Ercoli | Erica Arcangeli | Cristina Berardi | Lucilla Giulietti | Claudia Boccardi	present	urn:lsid:marinespecies.org:taxname:137205	Caretta caretta	Animalia	Linnaeus, 1758

extendedMeasurementOrFact (eMoF) extension:

The extendedMeasurementOrFact extension (eMoF) for survey or sighting data contains additional attributes and measurements recorded during the survey, such as those regarding the Research Vessel, environmental conditions, and/ or animal measurements. These attributes are linked to the Occurrence extension using the `occurrenceID`. The example dataset contain measurements regarding the sampling method; speed and height of the Research Vessel as platform; wind force; sighting distance; as well as the count and developmental stage of the biological entity.

id	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementUnit	measurementUnitID
TURTLE_CBAR_0001	TURTLE_CBAR_0001	BEAUFORT WIND FORCE	http://vocab.nerc.ac.uk/collection/P01/current/WMOCWFBF	0	Beaufort scale	
TURTLE_CBAR_0001	TURTLE_CBAR_0001	Platform height	http://vocab.nerc.ac.uk/collection/P01/current/AHSLZZ01	29	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA/
TURTLE_CBAR_0001	TURTLE_CBAR_0001	Sampling method	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100003	visual observation from ferries		
TURTLE_CBAR_0001	TURTLE_CBAR_0001	Speed of measurement platform relative to ground surface {speed over ground} by unspecified GPS system	http://vocab.nerc.ac.uk/collection/P01/current/APSAGP01	23.291	Knots (nautical miles per hour)	http://vocab.nerc.ac.uk/collection/P06/current/UKNT/
TURTLE_CBAR_0005	TURTLE_CBAR_0005	Developmental stage of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01			
TURTLE_CBAR_0005	TURTLE_CBAR_0005	Count (of assayed sample) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01	20	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA/
TURTLE_CBAR_0005	TURTLE_CBAR_0005	Sighting distance				

In addition to the measurements recorded by the example dataset, other measurements are also possible depending on the scope and aims of the survey project. The example dataset Incidental sea snake and turtle bycatch records from the RV Southern Surveyor voyage SS199510, Gulf of Carpentaria, Australia (Nov 1995) for example, contain information regarding the length and weight of the biological entity as follows:

extendedMeasurementOrFact (eMoF) extension:

id	measurementID	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementUnit	measurementUnitID
SS199510-001	SS199510-001-length	SS199510-001	Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX	1250	Millimetres	http://vocab.nerc.ac.uk/collection/P06/current/UXMM
SS199510-001	SS199510-001-weight	SS199510-001	Weight	http://vocab.nerc.ac.uk/collection/P01/current/SPWGXX01	800	Grams	http://vocab.nerc.ac.uk/collection/P06/current/UGRM
SS199510-002	SS199510-002-length	SS199510-002	Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX	1630	Millimetres	http://vocab.nerc.ac.uk/collection/P06/current/UXMM

id	measurementID	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementUnit	measurementUnitID
SS199510-002	SS199510-002-weight	SS199510-002	Weight	http://vocab.nerc.ac.uk/collection/P01/current/SPWGXX01	1477.7	Grams	http://vocab.nerc.ac.uk/collection/P06/current/UGRM

4.10.0.9 Microbes biomass and diversity

(example coming soon)

4.10.0.10 Phytoplankton biomass and diversity

This example deals with encoding phytoplankton observation data, including environmental data, into Darwin Core. Extracts from the actual data set LifeWatch observatory data: phytoplankton observations by imaging flow cytometry (FlowCam) in the Belgian Part of the North Sea, are used as an example. Phytoplankton data should adhere to guidelines by Martin-Cabrera et al., 2022: Best practices and recommendations for plankton imaging data management: Ensuring effective data flow towards European data infrastructures. Version 1.. Specific guidelines for the Occurrence and eMoF tables are documented in this publication, and we emphasize important fields in this example.

Event core:

The Event core contains events at the different levels and are linked together with `eventID` and `parentEventID`. In this example,the dataset contains records pointing to the origin, the in-situ sampling position as well as a record referring to the ex-situ collection of living specimens. In this case, the the event type information is provided in `type`. The recommended practice for providing the countryCode is to use an ISO 3166-1-alpha-2 country code. If additional information regarding licencing is provided, these can be populated under `rightsHolder` and `accessRights`. The remaining Event core fields provide location data including `datasetID` and `datasetName`, `locationID`, `waterBody`, `maximumDepthInMeters`, `minimumDepthInMeters`, `decimalLongitude`, `decimalLatitude`, `coordinateUncertaintyInMeters`, `geodeticDatum` and `footprintSRS`.

eventID	parentEventID	type	eventDate	modified	datasetID	datasetName	locationID	waterBody	country	countryCode	minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	geodeticDatum	coordinateUncertaintyInMeters	footprintSRS
TripNR3242		cruise	2017-05-05T13:18:00+00:00/2017-05-05T22:14:00+00:00	2021-10-21T15:52:00	https://marineinfo.org/id/dataset/4688	LifeWatch observatory data: phytoplankton observations...		North Sea	Belgium	BE					EPSG:4326		EPSG:4326
TripNR3242	TripNR3242		2017-05-08T20:44:00+00:00/2017-05-08T20:55:00+00:00	2021-10-21T15:52:00	https://marineinfo.org/id/dataset/4688		JN17_5			0	30.22	51.0131	1.90562	EPSG:4326		EPSG:4326
TripNR3242	TripNR3242		2017-05-08T20:50:00+00:00	2021-10-21T15:52:00	https://marineinfo.org/id/dataset/4688		JN17_5			0	1	51.01203	1.90217	EPSG:4326		EPSG:4326
TripNR3242	TripNR3242		2017-05-08T20:50:00+00:00	2021-10-21T15:52:00	https://marineinfo.org/id/dataset/4688		JN17_5			3	3	51.01203	1.90217	EPSG:4326		EPSG:4326

Occurrence extension:

The Occurrence extension contains data of each occurrence with an `occurrenceID` and is linked to the Event core with the `eventID`. The Occurrence extension should provide information on the `basisOfRecord` and `occurrenceStatus`. Scientific names and links to the World Register of Marine Species should be provided under `scientificName` and `scientificNameID`, respectively. Recommended best practice is to always use the term

of `MachineObservation` for imaging datasets derived from imaging instruments like this example, Additionally, the fields `identifiedBy` and `identificationVerificationStatus` are crucial to indicate whether the data has been validated, and by whom. While not shown in this example, `identificationReferences` and `associatedMedia` are also recommended to document the citation or algorithm software used in identification, and the persistent URL of annotated images, respectively. We know this dataset was validated because of information provided in the abstract and the eMoF table below, "identified and counted by image analysis and normalised to a unit volume of water body, validated by human".

eventID	occurrenceID	modified	basisOfRecord	occurrenceStatus	scientificNameID	scientificName	identifiedBy	identificationVerificationSt
TripNR3242TripStationNR16781Midas2021	TripNR3242TripStationNR16781Midas2021:Actinoptychus_senarius_5	2021-10-21	MachineObservation	present	urn:lsid:marinespecies.org:taxname:148948	Actinoptychus senarius	Flanders Marine Institute	ValidatedByHuman
TripNR3242TripStationNR16781Midas2021	TripNR3242TripStationNR16781Midas2021:Actinoptychus_splendens_5	2021-10-21	MachineObservation	present	urn:lsid:marinespecies.org:taxname:148949	Actinoptychus splendens	Flanders Marine Institute	ValidatedByHuman
TripNR3242TripStationNR16781Midas2021	TripNR3242TripStationNR16781Midas2021:Actinoptychus_sp_5	2021-10-21	MachineObservation	present	urn:lsid:marinespecies.org:taxname:148947	Actinoptychus	Flanders Marine Institute	ValidatedByHuman

extendedMeasurementOrFact (eMoF) extension:

The eMoF extension contains the environmental and measurement information and data of each occurrence. This extension is also linked to the Event core using the `eventID`, and linked to the Occurrence extension table using the `occurrenceID`. The various measurements are populated with `measurementID`, `measurementType`, `measurementTypeID`, `measurementUnit`, `measurementUnitID`, `measurementValue`, `measurementValueID`, `measurementAccuracy`, `measurementMethod`, `measurementDeterminedBy` and `measurementDeterminedDate`. In the example dataset, the LifeWatch observatory data was compiled using imaging flow cytometry (FlowCam) to observe and identify phytoplankton in the Belgian Part of the North Sea and recorded a number of measurements including abundance, lifestages, sampling device information as well as environmental measurements such as water temperature, salinity and conductivity with accompanying vocabulary.

id	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID	measurementDeterminedBy	measurementMethod
TripNR3242		Platform Name	http://vocab.nerc.ac.uk/collection/Q01/current/Q01000001/	Simon Stevin	http://vocab.nerc.ac.uk/collection/C17/current/11SS/			Flanders Marine Institute	
TripNR3242TripStationNR16781Midas2021	TripNR3242TripStationNR16781Midas2021:Actinoptychus_sp_5	Abundance of biological entity specified elsewhere per unit volume of the water body	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/	0.528	urn:lsid:marinespecies.org:taxname:148947	Actinoptychus sp	http://vocab.nerc.ac.uk/collection/P06/current/UCPL	Flanders Marine Institute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human
TripNR3242TripStationNR16781Midas2021	TripNR3242TripStationNR16781Midas2021:Actinoptychus_senarius_5	Abundance of biological entity specified elsewhere per unit volume of the water body	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/	0.528	urn:lsid:marinespecies.org:taxname:148948	Actinoptychus senarius	http://vocab.nerc.ac.uk/collection/P06/current/UCPL	Flanders Marine Institute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human
TripNR3242TripStationNR16781Midas2021	TripNR3242TripStationNR16781Midas2021:(Pseudo-)pediastrum_5	Abundance of biological entity specified elsewhere per unit volume of the water body	http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01/	0.528	urn:lsid:marinespecies.org:taxname:160560	(Pseudo-)	http://vocab.nerc.ac.uk/collection/S11/current/S1116/	Flanders Marine Institute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human

id	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID	measurementDeterminedBy	measurementMethod
TripNR3242TripStationNR16781MidSamTripActionID105598Occurrence1DTA__105598__Actinoptychus senarius__5			http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01/		http://vocab.nerc.ac.uk/collection/S11/current/S1116/				
TripNR3242TripStationNR16781MidSamTripActionID105598		sampling aperture diameter	http://vocab.nerc.ac.uk/collection/Q01/current/Q01000012/	0.4		meter	http://vocab.nerc.ac.uk/collection/P06/current/ULAA/	Flanders Marine Institute	
TripNR3242TripStationNR16781MidSamTripActionID105598		sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q01000002/	Planktonnet Apstein	http://vocab.nerc.ac.uk/collection/L22/current/TOOL0978/			Flanders Marine Institute	
TripNR3242TripStationNR16781MidSamTripActionID105598		sampling mesh size	http://vocab.nerc.ac.uk/collection/Q01/current/Q01000015/	55		micrometer	http://vocab.nerc.ac.uk/collection/P06/current/UMIC/	Flanders Marine Institute	
TripNR3242TripStationNR16781MidSamTripActionID105598UW		Conductivity of the water body	http://vocab.nerc.ac.uk/collection/P01/current/CNDCZZ01/	3.916		Siemens per metre	http://vocab.nerc.ac.uk/collection/P06/current/UECA	Flanders Marine Institute	Electrical conductivity of the water body by thermos-alinograph, based on the Under-WaySystem of the ship
TripNR3242TripStationNR16781MidSamTripActionID105598UW		Practical salinity of the water body	http://vocab.nerc.ac.uk/collection/P01/current/PSALPR01/	34.295		Grams per kilogram	http://vocab.nerc.ac.uk/collection/P06/current/UGKG/	Flanders Marine Institute	Practical salinity of the water body based on water from the Under-WaySystem of the ship
TripNR3242TripStationNR16781MidSamTripActionID105598UW		Temperature of the water body	http://vocab.nerc.ac.uk/collection/P01/current/TEMPPR01/	11.881		Degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA/	Flanders Marine Institute	Temperature of the water body based on water from the Under-WaySystem of the ship

4.10.0.11 Seagrass cover and composition

The structure of the Event, Occurrence and extendedMeasurementOrFact extensions for Seagrass Cover & Composition is based on community feedback organised through the the Scientific Committee on Oceanic Research (SCOR): Coordinated Global Research Assessment of Seagrass System (C-GRASS). We acknowledge the work that the C-grass SCOR work group has done to develop a proposed scheme for completing Seagrass related extension files.

Here encode seagrass survey data into Darwin Core according to the ENV-DATA approach and using sections of the actual data set of Seagrass Monitoring at Chengue Bay, Colombia as an example dataset.

Event core:

The Event core table is created by extracting all events and attributes. All events are linked together using `eventID` and `parentEventID`. `eventDate` is populated at the transect level with the recommended format that conforms to ISO 8601-1:2019. `habitat` is populated as a category or description of the habitat in which the event occurred. Additional `fieldNotes` can also be provided if applicable. The recommended best practice for `countryCode` is to use an ISO 3166-1-alpha-2 country code. The remaining Event core fields comprise of location data including `maximumDepthInMeters`, `minimumDepthInMeters`, `decimalLongitude`, `decimalLatitude`, `coordinateUncertaintyInMeters`, `footprintWKT` and `footprintSRS`. Addtionally in the Event core, it is

recommended to further include information regarding `license`, `rightsHolder`, `bibliographicCitation`, `institutionID`, `datasetID`, `institutionCode` and `datasetName`.

eventID	parentEventID	eventDate	habitat	fieldNotes	countryCode	minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	coordinateUncertaintyInMeters	footprintWKT	footprintSRS
USBsg-chengue-pastocoral		2019-05-13	seagrass	no notes	CO	0.8	2	11.32021806	-74.12753684	10	POLYGON ((-74.1273259763024 11.320475512862,-74.1272978004008 11.3201655779439))	EPSG:4326
USBsg-chengue-pastomanglar		2019-05-14	seagrass	no notes	CO	0.8	0.8	11.31977189	-74.12536879	10	POLYGON ((-74.1253370891273 11.3195001294432,-74.1253333743154 11.3194968146313))	EPSG:4326
USBsg-chengue-pastocoral-SquidPopTransect1	USBsg-chengue-pastocoral	2019-05-13	seagrass	no notes	CO	0.8	2	11.32039927	-74.12737404	50	POINT (-74.1273740410759 11.3203992721869)	EPSG:4326
USBsg-chengue-pastocoral-SquidPopTransect2	USBsg-chengue-pastocoral	2019-05-13	seagrass	no notes	CO	0.8	2	11.32027662	-74.1273989	50	POINT (-74.1273989021655 11.3202766241445)	EPSG:4326

Occurrence extension:

The Occurrence extension table contain data for each occurrence with an `occurrenceID` and is linked to the Event core with the `eventID`. This table should provide information on the `basisOfRecord` and `occurrenceStatus`. Scientific names and links to the World Register of Marine Species should be provided under `scientificName` and `scientificNameID`, respectively. If a species was identified by an expert, the field `identifiedBy` can be populated. If the species is well-known by another common name, this name can be provided under `vernacularName`.

eventID	occurrenceID	basisOfRecord	occurrenceStatus	scientificNameID	scientificName
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	HumanObservation	present	urn:lsid:marinespecies.org:taxname:374720	Thalassia testudinum
USBsg-chengue-pastomanglar	USBsg-chengue-manglar-tt	HumanObservation	present	urn:lsid:marinespecies.org:taxname:374720	Thalassia testudinum
USBsg-chengue-pastocoral-SquidPopTransect1	USBsg-chengue-pastocoral-fish-001	HumanObservation	present	urn:lsid:marinespecies.org:taxname:158815	Halichoeres bivittatus
USBsg-chengue-pastocoral-SquidPopTransect1	USBsg-chengue-pastocoral-fish-002	HumanObservation	present	urn:lsid:marinespecies.org:taxname:158932	Lactophrys triqueter

extendedMeasurementOrFact (eMoF) extension:

The eMoF table contains the measurement information and data of each occurrence. This extension is also linked to the Event core using the `eventID`, and linked to the Occurrence table using the `occurrenceID`. The various measurements are populated with `measurementType`, `measurementTypeID`, `measurementUnit`, `measurementUnitID`, `measurementValue`, `measurementValueID`, `measurementAccuracy`, `measurementMethod`, `measurementDeterminedBy` and `measurementDeterminedDate`. The example dataset of Seagrass Monitoring at Chengue Bay, Colombia recorded a number of measurements and can be used as an example of how to populate the respective fields:

eventID	occurrenceID	measurementID	measurementType	measurementTypeID	measurementValue	measurementUnit	measurementUnitID
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-PhyQ01	WaterTemp	http://vocab.nerc.ac.uk/collection/P01/current/TEMPMP01/	29.23	Degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-PhyQ02	Salinity	http://vocab.nerc.ac.uk/collection/P01/current/SSALSL01/	36	Parts per thousand	http://vocab.nerc.ac.uk/collection/P06/current/UPPT/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-PhyQ03	Dissolved oxygen	http://vocab.nerc.ac.uk/collection/P01/current/DOXYSE02/	6.58	Milligrams per litre	http://vocab.nerc.ac.uk/collection/P06/current/UMGL/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1C1-shoot-01	Shoot Density	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL02/	128	Number per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UPMS/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1C1-leafLenght-01	Leaf Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSMAXLX/	18	Centimetres	http://vocab.nerc.ac.uk/collection/P06/current/ULCM/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-DryBiomass	Total Dry Biomass	http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.32055	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-biomassGL	Dry biomass of green leaves	http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.05575	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-biomassNGL	Dry biomass of non green leaves	http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.1469	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/

eventID	occurrenceID	measurementID	measurementType	measurementTypeID	measurementValue	measurementUnit	measurementUnitID
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-biomassSH	Dry biomass of the shoots	http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.07625	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N2-biomassR	Dry biomass of the roots	http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.0385	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N2-biomassRIZ	Dry biomass of the rizome	http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.02725	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N2-biomassOTH	Dry biomass of other seagrass species	http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/

4.10.0.12 Zooplankton biomass and diversity

Here we will encode zooplankton observation and environmental data into Darwin Core. Extracts from the actual dataset LifeWatch observatory data: zooplankton observations by imaging (ZooScan) in the Belgian Part of the North Sea, are used as an example. As with the phytoplankton example, zooplankton data should also adhere to guidelines by Martin-Cabrera et al., 2022: Best practices and recommendations for plankton imaging data management: Ensuring effective data flow towards European data infrastructures. Version 1.. These Best Practices indicate that for imaging data like this example, the fields `identificationVerificationStatus` and `identifiedBy` are crucial to know whether the data has been validated, and by whom. `identificationReferences` is used to document the citation or algorithm software used in identification. When possible, `associatedMedia` should also be populated with the persistent URL of annotated images.

Event core:

The Event core contains events at the different levels and are linked together with `eventID` and `parentEventID`. In this example,the dataset contains records pointing to the origin, the in-situ sampling position as well as a record referring to the ex-situ collection of living specimens. In this case, the the event type information is provided in `type`. The recommended practice for providing the countryCode is to use an ISO 3166-1-alpha-2 country code. If additional information regarding licencing is provided, these can be populated under `rightsHolder` and `accessRights`. The remaining Event core fields provide location data including `datasetID` and `datasetName`, `locationID`, `waterBody`, `maximumDepthInMeters`, `minimumDepthInMeters`, `decimalLongitude`, `decimalLatitude`, `coordinateUncertaintyInMeters`, `geodeticDatum` and `footprintSRS`.

eventID	parentEventID	type	eventRemarks	eventDate	modified	datasetID	datasetName	locationID	waterBody	country	minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	geodeticDatum	footprintSRS
TripNR2547		cruise		2013-07-22T06:58:00+00:00/2013-07-22T16:58:00+00:00	2021-06-23T00:00:00+00:00	https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...		Belgian Part of the North Sea	Belgium					EPSG:4326	EPSG:4326
TripNR2547	TripNR2547:Haul:978	Visit		2013-07-22T07:13:00+00:00/2013-07-22T07:26:00+00:00	2021-06-23T00:00:00+00:00	https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...	130			0	13.4	51.27083333	2.905	EPSG:4326	EPSG:4326
TripNR2547	TripNR2547:Haul:975	Preservation		2013-07-22T07:22:00+00:00	2021-06-23T00:00:00+00:00	https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...	130			0	0	51.26873182	2.901797	EPSG:4326	EPSG:4326

eventID	parentEventID	eventRemarks	eventDate	modified	datasetID	datasetName	locationID	waterBody	country	minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	footprintSRS
TripNR2547	TripNR2547	TripStation	NR1978	2022-07-22T07:22:00+00:00	2022-07-23	https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...	130		3	3	51.26873	2.901797	EPSG:4326

Occurrence extension:

The Occurrence extension contains data of each occurrence with an `occurrenceID` and is linked to the Event core with the `eventID`. The Occurrence extension should provide information on the `basisOfRecord` and `occurrenceStatus`. Scientific names and links to the World Register of Marine Species should be provided under `scientificName` and `scientificNameID`, respectively.

eventID	occurrenceID	modified	basisOfRecord	occurrenceStatus	scientificNameID	scientificName	identifiedBy	identificationVerificationStatus	References
TripNR2547	TripStation	NR1978	2547	TripStation	2022-07-22	MidwaterTrawl	HumanObservation	2022-07-22T00:00:00+00:00	occurrenceID	TA23024	https://marinespecies.org:taxname:1135	Amphipoda	Jonas Mortelmans	130	ValidatedByHuman	RandomForest
TripNR2547	TripStation	NR1978	2547	TripStation	2022-07-22	MidwaterTrawl	HumanObservation	2022-07-22T00:00:00+00:00	occurrenceID	TA23024	https://marinespecies.org:taxname:882	Annelida	Jonas Mortelmans	130	ValidatedByHuman	RandomForest
TripNR2547	TripStation	NR1978	2547	TripStation	2022-07-22	MidwaterTrawl	HumanObservation	2022-07-22T00:00:00+00:00	occurrenceID	TA23024	https://marinespecies.org:taxname:106671	Anomura	Jonas Mortelmans	130	ValidatedByHuman	RandomForest
TripNR2547	TripStation	NR1978	2547	TripStation	2022-07-22	MidwaterTrawl	HumanObservation	2022-07-22T00:00:00+00:00	occurrenceID	TA23024	https://marinespecies.org:taxname:146421	Appendicularia	Jonas Mortelmans	130	ValidatedByHuman	RandomForest

extendedMeasurementOrFact (eMoF) extension:

The eMoF extension table contains the measurement information and data of each occurrence. This extension is also linked to the Event core using the `eventID`, and linked to the Occurrence table using the `occurrenceID`. The various measurements are populated with `measurementType`, `measurementTypeID`, `measurementUnit`, `measurementUnitID`, `measurementValue`, `measurementValueID`, `measurementAccuracy`, `measurementMethod`, `measurementDeterminedBy` and `measurementDeterminedDate`. The example dataset of LifeWatch observatory data: zooplankton observations by imaging (ZooScan) in the Belgian Part of the North Sea recorded some ENV-DATA and organism measurements the can be used as an example of how to populate the respective fields, including conductivity of the water body; concentration of chlorophyll-a per unit volume of the water body; sampling instrument name; sampling net mesh size; lifestage of the organism observed; and abundance of the organism observed.

id	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID	measurementDeterminedBy	measurementMethod
TripNR3256	TripStation	NR1715	MidwaterTrawl	sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Planktonnet WP2	<http://vocab.nerc.ac.uk/collection/L22/current/TOOL0979/>				
TripNR3256	TripStation	NR1715	MidwaterTrawl	sampling net mesh size	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100015/	200		micrometer	http://vocab.nerc.ac.uk/collection/P06/current/UMIC/		
TripNR3529	TripStation	NR1924	MidwaterTrawl	conductivity of the water body	http://vocab.nerc.ac.uk/collection/P01/current/CNDCZZ01/	4.05		Siemens per metre	http://vocab.nerc.ac.uk/collection/P06/current/UECA/	Flanders Marine Institute	Electrical conductivity of the water body by thermosalinograph, based on the UnderWaySystem of the ship

id	occurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID	measurementDeterminedBy	measurementMethod
TripNR3529	TripStationNR19243	Mass Concentration of chlorophyll-a per unit volume of the water body	http://vocab.nerc.ac.uk/collection/P01/current/CPHLHPP1/	1.42		Micrograms per litre	http://vocab.nerc.ac.uk/collection/P06/current/UGPL/	Flanders Marine Institute	Concentration of chlorophyll-a per unit volume of the water body [particulate >GF/F phase] by filtration, acetone extraction and high performance liquid chromatography (HPLC)
TripNR2547	TripStationNR23024	Abundance of biological entity specified elsewhere per unit volume of the water body	http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01/	1	http://vocab.nerc.ac.uk/collection/S11/current/S1152/	Annelida__sub2__130		Flanders Marine Institute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human
TripNR2547	TripStationNR23024	Abundance of biological entity specified elsewhere per unit volume of the water body	http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/	0.5	occurenceIDTA23024__Annelida__sub2__130	individuals per cubic metre	http://vocab.nerc.ac.uk/collection/P06/current/UPMM/	Flanders Marine Insitute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human

4.10.1 Other data types

Content

- Multimedia data
- Habitat data
- Tracking data

4.10.1.1 Multimedia data (Acoustic, Imaging)

If you have multimedia data (e.g. images, acoustic, video) that you want to publish alongside your dataset, you can do so by documenting information in the `associatedMedia` field in your Occurrence table. The usage of this field requires the media in question to be hosted somewhere with a persistent URL of the annotated image(s), e.g., a publication, museum database, etc. Then you simply copy this link to the `associatedMedia` field for a given occurrence. You may also include a concatenated list if you need to list multiple sources.

While there are Core types and extensions (e.g., Audubon Core and Simple Multimedia extension) designed for image, video, and audio files, these data file types are not currently processed by OBIS. Thus for now we recommended to include links in the `associatedMedia` field. Stay tuned however, as OBIS is looking to incorporate the Simple Multimedia extension.

Martin-Cabrera et al., 2022 have produced a best practices for datasets with plankton imaging data that can also apply to acoustic and other imaging data types. Following their guidelines, we strongly recommend including the following terms in your Occurrence table for either of these data types:

- `basisOfRecord` - recommended best practice is to always use the term of `MachineObservation`, especially for imaging datasets derived from imaging instruments
- `identifiedBy` - name(s) of persons involved in verifying taxon identification, particularly if automatic identification was made by a software and then validated by a human
- `identificationVerificationStatus` - categorical indicator for the extent of taxonomic identification verification. Recommended to use PredictedByMachine or ValidatedByHuman
- `identificationReferences` - references used in identification (e.g. citation and version of software or algorithm that identified taxa)

The fields `identifiedBy` and `identificationVerificationStatus` are crucial to indicate whether an observation has been validated, and by whom. These fields allow users to filter data when `basisOfRecord` = MachineObservation, so that they can be confident in the taxonomic identification when `identificationVerificationStatus` = ValidatedByHuman (Martin-Cabrera et al., 2022).

The `identificationVerificationStatus` also has implications for documenting grouped occurrences, particularly for planktonic organisms. For example, if all identifications for a specific taxon in a sample has the same `identificationVerificationStatus`, you ony need **one** occurrence record with one associated unique occurrenceID. Then, the summed count or concentration for that taxon can be reported in the eMoF as, e.g. "Abundance of biological entity specified elsewhere per unit volume of the water body". However, if individuals of a taxon have more than one `identificationVerificationStatus` (e.g. ValidatedByHuman and PredictedByMachine), you will need **two** occurrence records with associated unique occurrenceIDs. The two records will document the same taxon with different `identificationVericationStatus`, and with different summed concentrations of abundance reported in the eMoF.

Example Resources: Martin-Cabrera et al. (2022) have created a best practices document for plankton imaging data that you can reference. To see an example imaging dataset implementing these best practices, see the supplementary material of Establishing Plankton Imagery Dataflows Towards International Biodiversity Data Aggregators.

Data originating from ROV (Remote Operating Vehicle) observations may require additional processing. Ocean Networks Canada (ONC) is developing a pipeline for publishing ROV data to OBIS. ROV datasets should have:

- An Event core that documents the hierarchical nature of ROV dives (e.g., ROV dives nested within a cruise)
- Occurrence and eMoF extensions to record taxonomic and other measurement data e.g., from sensors.

ONC's pipeline outlines the importance of including `identifiedBy` in order to vet taxon identifications by experts.

4.10.1.2 Habitat data

Event Core is perfect for enriching OBIS with interpreted information such as biological community, biotope or habitat type (collectively referred to as 'habitats'). However, the unconstrained nature of the terms `measurementTypeID`, `measurementValueID`, and `measurementUnitID` leads to a risk that habitats measurements are structured inconsistently within the Darwin Core Archive standard and as a result, are not easily discoverable, understood or usable.

As a result, members of the European Marine Observation and Data Network (EMODnet) Seabed Habitats and Biology thematic groups have produced a technical report Duncan et al. (2021) that provides guidance on using the Darwin Core eMoF extension to submit habitat data to OBIS, following the ENV-DATA approach and using Seabed Habitats as a use case. Note that the guidelines and structuring approach outlined in this document has not yet been approved or accepted at the global level and is only a recommended approach as agreed upon by EMODnet Seabed Habitats, EMODnet Biology, and OBIS. The implementation at the EurOBIS level may be considered a pilot.

The overarching principles are summarised here. Note that because of the numerous classification systems and priority habitat lists in existence, it is not possible to point to a single vocabulary for populating each

of `measurementTypeID`, `measurementValueID` and `measurementUnitID`, as for other measurement types, so below are the *types* of information to include, with an example, as recommended by Duncan et al. (2021):

- `measurementTypeID`: A machine-readable URI or DOI reference describing the (version of the) classification system itself. For example: https://dd.eionet.europa.eu/vocabulary/biodiversity/eunishabitats/
- `measurementValueID`: If available, a machine-readable URI describing the habitat class in "measurementValue". For example: https://dd.eionet.europa.eu/vocabulary/biodiversity/eunishabitats/A5.36
- `measurementUnitID`: null because habitat types are unitless.

Please consult the Duncan et al. (2021) technical report (title: A standard approach to structuring classified habitat data using the Darwin Core Extended Measurement or Fact Extension; note you must refine search to Technical Reports from 2021 to identify this report) for more details, including:

- how to handle a single event with multiple habitat measurements
- recommended vocabularies and terms for common habitat classification systems
- example eMoF table

For filling measurementType with habitat-related data and/or the dwc:habitat column, you should reference the NERC vocabulary search. While the Coastal and Marine Ecological Classification Standard (CMECS) and the Environment Ontology (ENVO) also contain habitat vocabularies, OBIS recommends the use of NERC vocabulary. If other vocabularies are used, please provide the NERC vocabulary equivalent as additional records in the eMoF table.

4.10.1.3 Tracking data

Encoding Tracking data into Darwin Core follows the same standards as that of survey/sighting data. Tracking data should additionally indicate the accuracy in latitudinal and longitudinal measurements received from the positioning system, grouped by location accuracy classes, recorded in the `coordinateUncertaintyInMeters` field. The Ocean Tracking Network (OTN) has developed some guidelines for formatting this type of data in Darwin Core. We summarize the main points below.

Using Event core for tracking data is recommended as there can be multiple events involved in tracking an organism. There are capture/tag and release events, receiver deployment events, and detection occurrences. Note that the capture and release of an organism are not considered to be distinct Occurrence records because they are not natural occurrences. Thus, in the Event core table you may record unique events for:

- The capture of an animal
- The release of an animal
- The deployment of a listening (or receiver) station

Information pertaining to a specific individual is linked by a unique `organismID`. You can use `eventID`s associated with a receiver to record detection occurrences in the Occurrence table. One organism may then have multiple occurrences (and thus multiple occurrenceIDs), but the same `organismID`. Any measurements for an organism taken during capture can be recorded in the extendedMeasurementsOrFact extension, linked to the core by the capture event's `eventID` as well as the unique `organismID`. For more details, see the DwC guidelines for biologging.

Extracts from the extendedMeasurementOrFact Extension (eMoF) of the actual dataset Ningaloo Outlook turtle tracking of Green turtles (Chelonia mydas), Western Australia (2018-present), are shown as an example tracking dataset, following ARGOS Location class codes.

extendedMeasurementOrFact (eMoF) extension:

id	measurementID	occurrenceID	measurementType	measurementValue	measurementValueID
2347540	2347540-argosclass	2347540	ARGOS Location Class	A	http://vocab.nerc.ac.uk/collection/R05/current/A
2347541	2347541-argosclass	2347541	ARGOS Location Class	B	http://vocab.nerc.ac.uk/collection/R05/current/B
2347542	2347542-argosclass	2347542	ARGOS Location Class	2	http://vocab.nerc.ac.uk/collection/R05/current/2
2347543	2347543-argosclass	2347543	ARGOS Location Class	3	http://vocab.nerc.ac.uk/collection/R05/current/3

Ensuring Data Quality

Chapter 5

Data quality control

OBIS ignores records that do not meet a number of standards. For example, all species names need to be matched against an authoritative taxonomic register, such as the World Register of Marine Species. In addition, quality is checked against the OBIS required fields as well as against any impossible values. OBIS checks, rejects and reports the data quality back to the OBIS nodes, but never change records. The OBIS tier 2 nodes are responsible for the data quality and communicate errors back to the data providers. A number of QC tools are developed to help data providers and OBIS nodes:

- QC tool for species names
- QC tool for geography and data format

For specific concerns regarding quality control checks or issues, please submit a GitHub ticket to the OBIS QC repository.

5.1 Why are records dropped?

Records can be dropped and therefore not published with your dataset for a number of reasons, including:

- The species is not marine
- The 'scientificName' or `scientificNameID` did not match with WoRMS
- Issues with coordinates:
  - No coordinates given
  - `decimalLatitude` or `decimalLongitude` out of range
- The coordinate is zero

For each dataset published, a quality report is generated where the number of dropped records and other quality issues will be flagged. Such reports can also be found when searching for data in OBIS. For example, if we searched for 'Crustacea' records, the following data quality report is given:

We can see that >110,222 Crustacean records have been dropped, mostly due to records missing coordinates or species being flagged as non-marine. Because species are determined as being marine by WoRMS, we acknowledge that sometimes species are marked as `not_marine` erroneously.For specific advice on this topic, see the common QC issues page.

To minimize the number of records dropped, be careful when formatting your data so that you are meeting the requirements.

Figure 5.1: Number of Crustacean records dropped

5.2 How to conduct Quality Control

Once you have formatted your data for OBIS, or have received a formatted dataset, it is important to run quality control checks before publishing the dataset on the IPT. The following is a list of various tools you can use to help you perform quality checks on your data:

- R package obistools
- EMODnet Biocheck
  - Web UI built on obistools. This tool requires your dataset to be published on an IPT (e.g., a test IPT such as https://ipt.gbif.org/ where your dataset will not be harvested by GBIF or OBIS). Note you are required to have a login to access an IPT
  - R package
- Lifewatch data services
- The US Integrated Ocean Observing System Standardizing Marine Bio Data Guide
- WoRMS taxon match tool
  - Other WoRMS web services, incl. taxon match
- Excel Conditional Formatting tool
  - Excel > Home > Conditional Formating > Highlight cells Rules > Duplicate values...
- GBIF data validator
- Python library for OBIS QC developed by Canadian Integrated Ocean Observing System
- R package and function Hmisc:: describe
  - Can give important summary statistics and identify numbers that don't match

5.2.1 Conducting QC with obistools

To use `obistools` to conduct quality control, you can follow this general order:

1. Check that the taxa match with WoRMS
   - `obistools::match_taxa`
2. Check that all required fields are present in the occurrence table
   - `obistools::check_fields`
3. Check coordinates

- Plot them on a map to identify any points that appear outside the scope of the dataset `obistools::plot_map`
- Identify points with obistools::identify_map
- Check that points are not on land `obistools::check_onland`
- Ensure depth ranges are valid `obistools::check_depth`

4. Check for statistical outliers which may have had data entry errors
   - obistools::check_outliers_species and obistools::check_outliers_dataset
5. Check that the eventID and parentEventID are structured correctly `obistools::check_eventids`
   - Ensure all eventIDs in extensions have matching eventIDs in the core table `obistools::check_extension_eventids`
6. Check that eventDate is formatted properly `obistools::check_eventdate`

5.2.2 QC with R package Hmisc

The R package Hmisc has the function `describe` which can help you identify any discrepancies in your dataset.

It will summarize each of your variables for a given data field. This can help you quickly identify any missing data and ensure the number of unique IDs is correct. For example, in an Occurrence table with 1000 records, there should be 1000 unique occurrenceIDs.

```
library(Hmisc)
library(Hmisc)
data<-read.csv("example_data_occur.csv")
describe(data)

 12  Variables      407  Observations
---------------------------------------------------------------------------------------------------------------
CollectionCode
       n  missing distinct    value
     407        0        1  BIOFUN1

Value      BIOFUN1
Frequency      407
Proportion       1
---------------------------------------------------------------------------------------------------------------
eventID
       n  missing distinct
     407        0       27

lowest : BIOFUN1_BF1A01 BIOFUN1_BF1A02 BIOFUN1_BF1A03 BIOFUN1_BF1A04 BIOFUN1_BF1A05
highest: BIOFUN1_BF1M3  BIOFUN1_BF1M4  BIOFUN1_BF1M6  BIOFUN1_BF1M8  BIOFUN1_BF1M9
---------------------------------------------------------------------------------------------------------------
occurrenceID
       n  missing distinct
     407        0      407

lowest : CSIC_BIOFUN1_1   CSIC_BIOFUN1_10  CSIC_BIOFUN1_100 CSIC_BIOFUN1_101 CSIC_BIOFUN1_102
highest: CSIC_BIOFUN1_95  CSIC_BIOFUN1_96  CSIC_BIOFUN1_97  CSIC_BIOFUN1_98  CSIC_BIOFUN1_99
```

This video shows how to use both `obistools` and `Hmisc` to conduct QC checks in R.

5.3 Data quality flags

As you are following the guidelines in this manual to format data, it is important to consider the potential quality flags that could be produced when your dataset is published to OBIS. OBIS performs a number of

automatic quality checks on the data it receives. This informs data users of any potential issues with a dataset they may be interested in. A complete list of flags can be found here but broadly speaking potential flags relate to issues with:

- Location - coordinates
- Event time - start, end dates
- Depth values out of range
- Taxonomic
- WoRMS name matching
- Non-marine or terrestrial

When you are filling these fields it is important to double check that correct coordinates are entered, eventDates are accurate and in the correct format, and depth values are in meters. Records may be rejected and not published with your dataset if the quality does not meet certain expectations. In other cases quality flags are attached to individual occurrence records.

We acknowledge that sometimes you may encounter a QC flag for data that is accurate. For example, you may document a depth value that gets flagged as `DEPTH_OUT_OF_RANGE`. Sometimes this occurs because your measured depth value is more accurate than the GEBCO bathymetry data which OBIS bases its depth data on. In these cases, you can ignore the flag, but we recommend adding a note in `eventRemarks`, `measurementRemarks`,or `occurrenceRemarks`.

The checks we perform as well as the associated flags are documented here.

5.3.1 QC Flags in downloaded data

There are several ways to inspect the quality flags associated with a specific dataset or any other subset of data. Data downloaded through the mapper and the R package will include a column named `flags` which contains a comma separated list of flags for each record. In addition, the data quality panel on the dataset and node pages has a flag icon which can be clicked to get an overview of all flags and the number of records affected.

This table includes quality flags, but also annotations from the WoRMS annotated names list. When OBIS receives a scientific name which cannot be matched with WoRMS automatically, it is sent to the WoRMS team. The WoRMS team will then annotate the name to indicate if and how the name can be fixed. Documentations about these annotations will be added here soon.

Clicking any of these flags will take you to a table showing the affected records. For example, this is a list of records from a single dataset which have the `no_match` flag, indicating that no LSID or an invalid LSID was provided, and the name could not be matched with WoRMS. The column `originalScientificName` contains the problematic names, as `scientificName` is used for the matched name.

At the top of the page there's a button to open the occurrence records in the mapper where they can be downloaded as CSV. The occurrence table also has the `flags` column, so when inspecting non matching names for example it's easy to check if the names at hand have any WoRMS annotations:

5.3.2 Inspecting QC flags with R

Inspecting flags using R is also very easy. The example below fetches the data from a single dataset, and lists the flags and the number of records affected. Notice that the `occurrence()` call has `dropped = TRUE` to make sure that any dropped records are included in the results:

```
library(robis)
library(tidyr)
library(dplyr)

# fetch all records for a dataset
```

```
df <- occurrence(datasetid = "f3d7798e-7bf2-4b85-8ed4-18f2c1849d7d", dropped = TRUE)

# unnest flags

df_long <- df %>%
  mutate(flags = strsplit(flags, ",")) %>%
  unnest(flags)

# get frequency per flag

data.frame(table(df_long$flags))
```

```
                              Var1 Freq
1                depth_exceeds_bath   78
2                  no_accepted_name   17
3                          no_depth    5
4                          no_match  138
5                       not_marine    2
6                          on_land    1
7      worms_annotation_await_editor    5
8  worms_annotation_reject_ambiguous    2
9    worms_annotation_reject_habitat    2
10             worms_annotation_todo    9
11     worms_annotation_unresolvable    7
```

This second example creates a list of annotated names for a dataset:

```
library(robis)
library(dplyr)
library(stringr)

# fetch all records for a dataset

df <- occurrence(datasetid = "f3d7798e-7bf2-4b85-8ed4-18f2c1849d7d", dropped = TRUE)

# only keep WoRMS annotations and summarize

df %>%
  select(originalScientificName, flags) %>%
  mutate(flags = strsplit(flags, ",")) %>%
  unnest(flags) %>%
  filter(str_detect(flags, "worms")) %>%
  group_by(originalScientificName, flags) %>%
  summarize(records = n())
```

```
   originalScientificName            flags                              records
   <chr>                             <chr>                                <int>
 1 Alcyonidium fruticosa             worms_annotation_reject_habitat          1
 2 Apicularia (Thapsiella) rudis sp. worms_annotation_unresolvable            1
 3 Arcoscalpellum vegae              worms_annotation_unresolvable            1
 4 Balanus evermanni                 worms_annotation_await_editor            1
 5 Chloramidae                       worms_annotation_reject_ambiguous        2
 6 Cleippides quadridentatus         worms_annotation_todo                    1
 7 Enhydrosoma hoplacantha           worms_annotation_reject_habitat          1
```

```
 8 Hippomedon setosa               worms_annotation_unresolvable          1
 9 Leionucula tenuis               worms_annotation_await_editor          1
10 Ophiocten borealis              worms_annotation_todo                  1
11 Ophiopholis gracilis            worms_annotation_todo                  1
12 Priapulus australis             worms_annotation_await_editor          1
13 Primnoella residaeformis        worms_annotation_unresolvable          1
14 Robulus orbigny                 worms_annotation_unresolvable          1
15 Tetraxonia                      worms_annotation_unresolvable          2
16 Tmetonyx barentsi               worms_annotation_await_editor          2
17 Triaxonida                      worms_annotation_todo                  6
```

5.4 How To Use MoF Report and Tool

A MEASUREMENT TYPES dataset report has been added regarding currently used measurementType and associated measurementTypeID(s), located near the bottom of the individual dataset pages (if measurementType in use for the dataset).

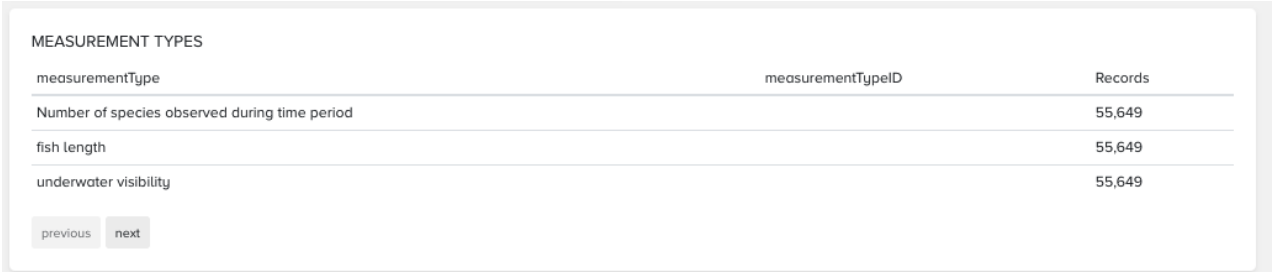
This new dataset report was derived from this MoF statistics report https://r.obis.org/mof/ and this active filtering MoF tool https://mof.obis.org/.

To more easily locate the datasets within your node that may have possible measurementType ID issues, use the MoF Statistics page: https://r.obis.org/mof/. This contains the list of Nodes currently using measurementType/measurementValue/measurementUnit with counts and percentage missing for the associated ID(s).

If there is a node in that list that you are interested in locating, searching for and possibly fixing MoF issues, select the Node from the list, then select a dataset (displaying a high percentage of missing ID(s)), and scroll down to the MEASUREMENT TYPE report

Example, selected OBIS USA,

then selected Florida Keys Reef Visual Census 1994, and scrolled down to MEASUREMENT TYPES section:



MEASUREMENT TYPES

measurementType	measurementTypeID	Records
Number of species observed during time period		55,649
fish length		55,649
underwater visibility		55,649

previous next

To locate other datasets using these MEASUREMENT TYPES, use this active filtering MoF tool https://mof.obis.org, sort by measurementType (click column header) and scroll to measurementType(s) of interest

For MEASUREMENT TYPE “Number of species observed during time period” has only one entry, which is missing associated ID. To see which datasets are using the listed measurementType, measurementTypeID combination, click on the number of records which is the last column.

All are from OBIS USA.

For MEASUREMENT TYPE “fish length” . . . To see which datasets are using this also listed measurementType, measurementTypeID combination, click on the number of records which is the last column.

There are two records for fish length, one missing an ID and the other using S06, which may not be the preferred ID for this measurementType:

first day of release			1,363
fish length	http://vocab.nerc.ac.uk/collection/S06/current/S0600101/	find	121,231
fish length			7,268,666
fish length size class median			238,425
Fish species is commonly caught			17

Also, while scrolling through this report, you may notice something you would like to further research, click the record count value to see a list of datasets and associated node(s) using this noted type/ID. NOTE: Current USE does not indicate CORRECT use:

To see BODC label for the provided ID, click the Find button, second last column:

This is showing a different label from the (variety of) measurementType provided.

To see which datasets are using a specific measurementType / ID combination, click the records count, last column:

Things you are looking to clean up:

- If measurementTypeID is empty this should be updated.
- If the same measurementType (with same meaning/purpose) is using multiple measurementTypeIDs, these should be fixed to a single, preferred BODC vocab value.

5.5 Geographic and data format quality control

These Data validation and QC services are available on the LifeWatch portal at http://www.lifewatch.be/data-services.

5.5.0.1 Geographical service

This service allows to upload a file and to plot the listed coordinates on a map. Using this web service does not require knowledge of GIS. This service allows a visual check of the available locations and makes it possible to easily identify points on land or outside the scope or study area. Geographic data are essential for OBIS and the experience is that a lot of these data is incomplete or contains errors. A visual check of the position of the sampling locations is thus a simple way of filtering out obvious errors and improving the data quality. Latitude and longitude need to be in WGS84, decimal degrees. This format is also necessary for the OBIS Schema and for uploading the dataset to IPT (Darwin Core).

5.5.0.2 OBIS data format validation

This is the most extensive check currently available and is available for data that are structured according to the OBIS Schema. This validation service checks the following items:

- Are all mandatory fields completed, what are the missing fields?
- Are the coordinates in the correct format (decimal degrees, taking into account the minimum and maximum possible values)?
- Are the sampling points on land or in water?
- Is the information in the date-fields valid (e.g. month between 1-12)?
- Can the taxon name be matched with WoRMS?

This tool undertakes several actions simultaneously. In a first step, this data service allows you to map your own column headers to the field names used in the OBIS Schema. When you then run the format validation service, the following actions are performed:

- A check of the mandatory fields of the OBIS Scheme. If mandatory fields would be missing, these will be listed separately, so you can complete them. Without these fields, the dataset cannot be accepted by the OBIS node.
- A listing of all the optional fields of the OBIS Scheme that are available in your file.
- Validation of the content of a number of fields:

- Latitude & longitude:
  - Are the values inside the world limit? (yes/no);
  - Are the values different from zero? (yes/no);
  - Are the values situated in the marine environment (sea/ocean) (=prerequisite of a marine dataset)? (yes/no)
- Date-related fields:
  - Do the year-month-day fields form a valid date? (yes/no)
  - Do the start- and end-date fields form a valid date? (yes/no)
- Scientific name:
  - Is the scientific name available in WoRMS? (yes/no)
  - When yes:
    * Indication whether taxon is marine or not
    * Indication whether taxon name is valid or not
    * Indication of the taxonomic rank

After matching with WoRMS, the report gives a brief overview containing:

- the number of exact matches
- the number of fuzzy (=non-exact) matches
- the number of non-matches
- the number of errors that might have occurred during matching

For each of the above steps, the result report lists the number of records that passes the check. The tool also makes a 'grand total' of these results, indicating if the quality of record is sufficient to be imported into OBIS, taking into account the results of the above mentioned checks.

If the file contains fields that do not match the OBIS schema, these are also listed. Fields that cannot be mapped to the OBIS schema will not be uploaded in OBIS.

After this data format check, a number of columns are added to the originally uploaded file, where the results of each step are listed. Each check is basically a yes/no question, which is translated to a 1 (yes) or 0 (no) value in the results file and is thus easy to interpret.

5.6 Common Quality Control issues

Content:

5.6.1 Uncertain temporal range

When the eventDate or temporal scope of your dataset is in question or provided in an invalid format (e.g., textual description), there are a number of options to ensure the most accurate date is provided.

1. You may provide a range of dates in the ISO 8601 format if the range of dates is certain. Do not include a date range if you are making assumptions. Notes about any assumptions or interpretations on date ranges can be documented in the `eventRemarks` field.
   - Be careful when entering date ranges. For example, entering 1870/1875-08-04 is equivalent to any date between 1870 and 1875-08-04. Date ranges can be used in this way to capture some level of uncertainty in when an event occurred.

2. If only parts of the date are known (e.g., year but not month and day), you may provide the date in ISO 8601 format while excluding the unknown elements. **Do not use zero** to populate incomplete dates, simply end the date with the known information (e.g., 2011-03 instead of 2011-03-00). Additionally, if the year is unknown, you should only populate the `month` and `day` fields because `eventDate` cannot be formatted to exclude year. In these cases, `eventDate` is not necessary to fill.
3. If date was provided as a textual description that is accurately interpretable, include the text description in the `verbatimEventDate` field. Then provide the interpreted date in ISO 8601 format in the `eventDate` field. Be sure to document any other important information in 'eventRemarks'.
4. For historical dates that do not conform to the ISO 8601 format, guidelines are still under development. But the dwc:class:GeologicalContext can be used to capture some information for records pertaining to fossilized specimens.

5.6.2 Uncertain geolocation

Sometimes locality information can be difficult to interpret, especially if records originate from historical data with vague descriptions, or descriptions/names of areas that no longer exist. If your dataset is missing `decimalLongitude` and `decimalLatitude`, but the locality name is given, there are a number of approaches you can take. You can:

- Use the OBIS Map Tool to obtain a WKT string for point, line, or polygon features to put in the `footprintWKT` field. The corresponding projection should be placed in the `footprintSRS` field. Note the accepted spatial reference system for OBIS is EPSG:4326 (WGS84). The Marine Regions Gazetteer is available for use within the Map tool to help find locations.
- Search for locations with the Marine Regions Gazetteer to obtain coordinates and a `locationID`. For information on how to use this tool, see below.
  - You can also use the Getty Thesaurus of Geographic Names or Google Maps. See below
  - **Note:** Always be sure to fill in the `georeferenceSources` field to indicate the sources you used to obtain locality information when appropriate
- If you have a set of points, a line, or a polygon (perhaps from the Map tool), you can find the centroid of the features using either obistools::calculate_centroid or PostGIS, and then enter this coordinate into the `decimalLatitude` and `decimalLongitude` fields. This PostGIS guideline will help you select a centroid that is guaranteed to fall within your designated area.
- Estimate `coordinateUncertaintyinMeters` that is wide enough to cover the area
  - If the location is provided as an array or WKT format, you can use R package obistools::calculate_centroid to obtain coordinate uncertainty.
  - Use the OBIS maptool to obtain from the "radius" column. This is only applicable for lines or polygons, not point features.

For data that only has **textual descriptions**:

- Try GBIF's GEOLocate Web Application. You can use this tool for one location at a time with the Standard Client, or upload a CSV file for batch processing. This tool lets you enter text descriptions in the "Locality String" field, and other relevant locality information (e.g. country, state, county) to obtain geographic coordinates.
- Use this Biodiversity Enhanced Location Services tool developed by VertNet. It can translate textual descriptions and provide decimalLatitude, decimalLongitude, geodeticDatum, and coordinateUncertaintyInMeters as a csv sent to an email address. For more information on this service, see the associated GitHub.

GBIF also provides some guidelines for difficult localities as well as other georeferencing tips for different geographic features, such as when only a distance or heading is provided (e.g., 10 km off Sao Paulo's coast, north of Fiji).

> Important note: If you are making any inferences and/or decisions about locality coordinates, please record this in the `georeferenceRemarks` field. Additional information about the local-

ity can also be stored in DwC terms such as `waterBody`, `islandGroup`, `island` and `country`. `locationAccordingTo` should provide the name of the gazetteer that was used to obtain the coordinates for the locality.

5.6.3 How to use OBIS Map Tool

A video tutorial on how to use our Map tool is available below. This video covers the following topics:

1. Estimating coordinates
2. Using the line and polygon tool
3. Obtaining and exporting WKT strings

Well-Known Text (WKT) strings are representations of the shape of the location and can be provided in the `footprintWKT` field. This is particularly useful for tracks, transects, tows, trawls, habitat extent, or when an exact location is not known. WKT strings can be created using the Map tool's WKT function. The Map tool also calculates a midpoint and a radius for line or polygon features, which can then be added to `decimalLongitude`, `decimalLatitude`, and `coordinateUncertaintyInMeters`, respectively. As mentioned above, the `obistools::calculate_centroid` function can be used to calculate the centroid and radius for WKT polygons. This wktmap tool can also be used to visualize and share WKT strings.

5.6.4 Using Getty Thesaurus & Google Maps to obtain locality coordinates

For both the Getty thesaurus and Google Maps you can simply search the name of a locality, for example the Cook Strait in New Zealand. The search result on the Getty thesaurus will bring you to a page where you can obtain `decimalLatitude` and `decimalLongitdue`.

For Google Maps, the coordinates can be found in the url after searching.

5.6.5 How to use Marine Regions Gazetteer tool

Marine Regions offers a marine gazetteer search engine to obtain geographic information and unique identifiers for marine regions. Once you have navigated to the gazetteer search engine, you have two options to search by: enter the name of the desired locality, or enter an MRGID code. Most likely you will have a locality name but not an MRGID. You may also select a placetype to search instead for types of regions that may be physical (e.g., seamount, bay, fjord, etc.) or administrative (e.g., exclusive economic zones, countries, etc). You can specify specific sources if known (e.g., published paper, organization, etc.). Finally, you can give a latitude/longitude coordinate with a radius around it to obtain a list of regions near that point.

For this example we will search by geographic name for the Bay of Fundy.

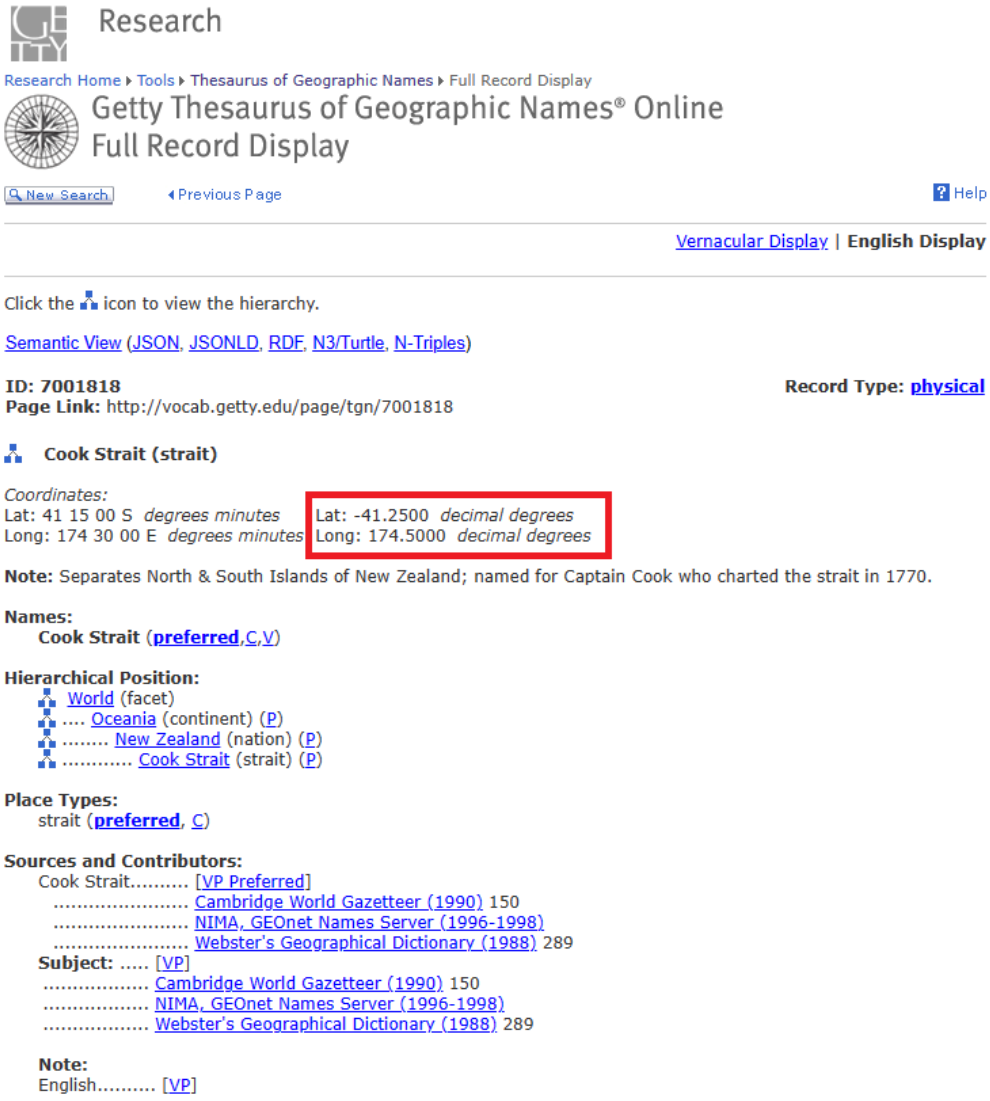


Figure 5.2: *Screenshot of Cook Strait page on the Getty Thersaurus*

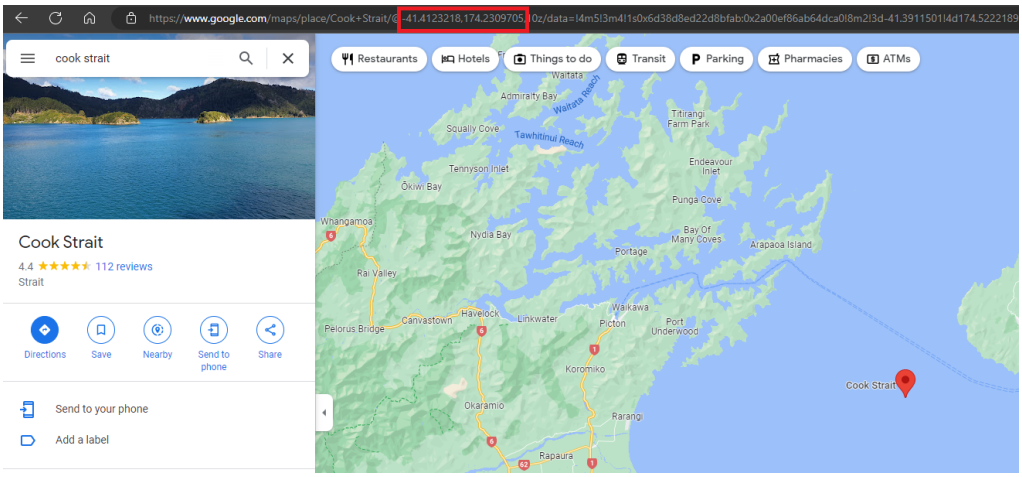
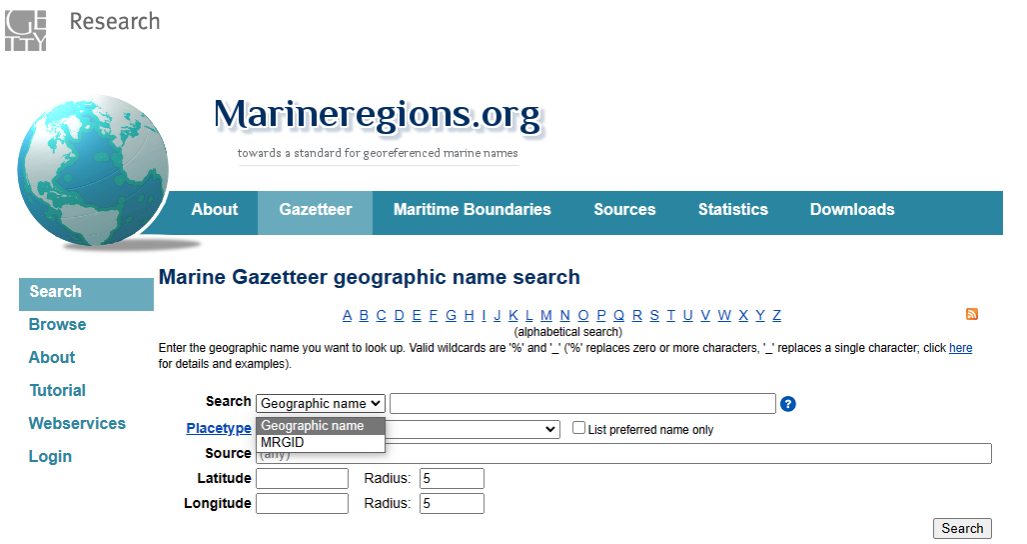
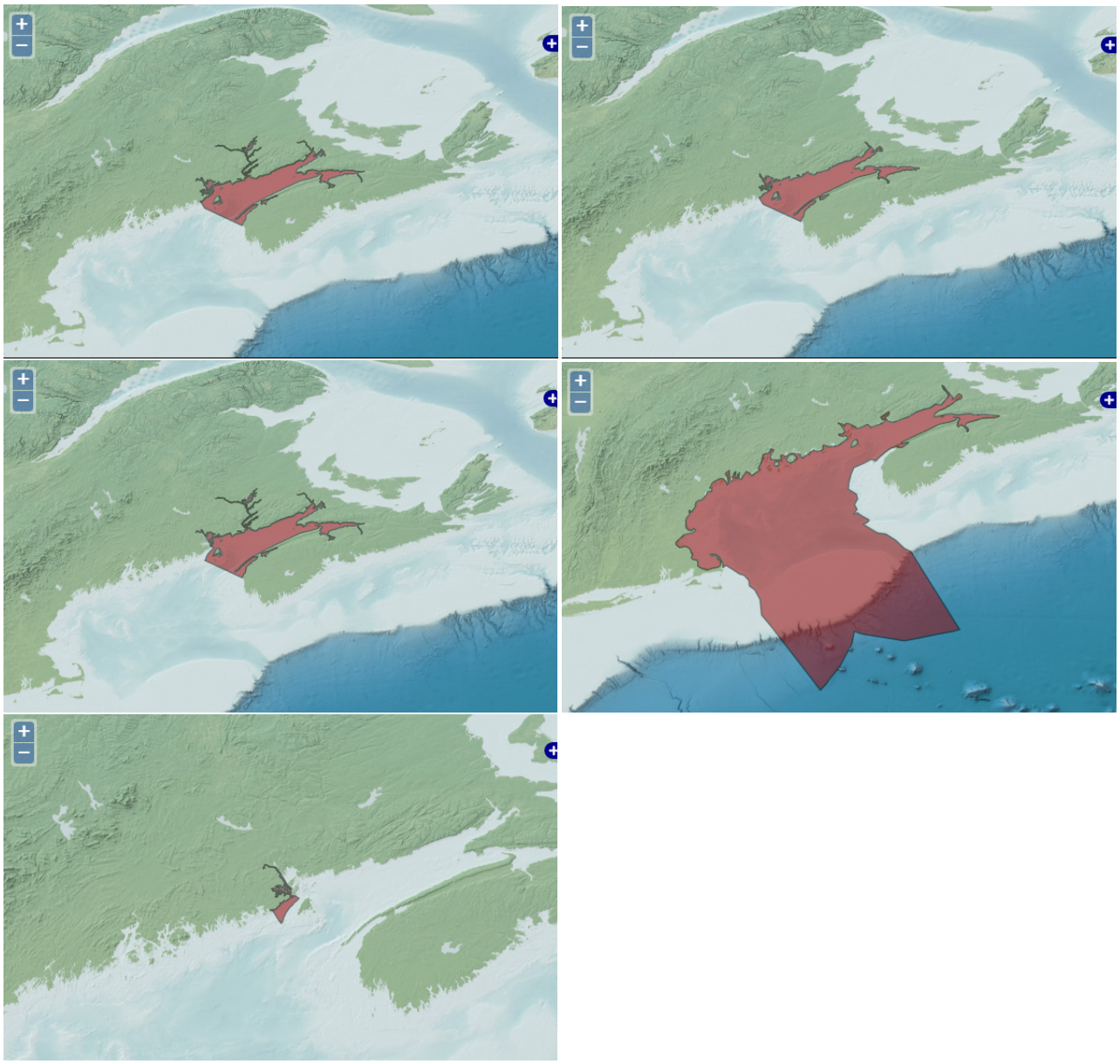


Figure 5.3: *Screenshot of Google Maps showing where coordinates can be found in the URL*



Our search returned 5 results from different sources (indicated in brackets). So how do we select the correct one? We can notice right away that the second result, from SeaVox SeaArea, has a preferred alternative, which when you click on the link brings you to the IHO Sea Area description for Bay of Fundy. So already we can likely drop SeaVox as a potential candidate. A good next step may be to compare the geographical extent for each to ensure it covers the desired area. If you are uncertain about exactly where your locality is, it may be better to be safe and choose a wider geographic region. Let's compare the maps for all 5 results:



Notice that no region has the exact same geographic extent. Let's select the IHO Bay of Fundy locality (the first search result) to ensure we are covering the entire area of the Bay of Fundy, but not the Gulf. Inspecting the rest of the page, there is a lot of other useful information we can use. We can populate the following OBIS fields for our dataset, copying the information outlined in the red boxes:

1. `locationID` from MRGID: http://marineregions.org/mrgid/4289
2. `decimalLatitude` and `decimalLongitude` latitude and longitude coordinates of the location's midpoint in decimal degrees: 44.97985204, -65.80601556
3. `coordinateUncertaintyInMeters` precision: 196726 meters

Since we are obtaining all this locality data from Marine Regions, we must also populate the `locationAccordingTo` field. Here, we will provide the name of the gazetteer we used to obtain the coordinates for the locality - in this case you would write "Marine Regions". In `georeferenceRemarks` we must document that the coordinates are the region's midpoint, that locality information was inferred by geographic name, and, where applicable, place the original locality name in the field `verbatimLocality`. Finally, the

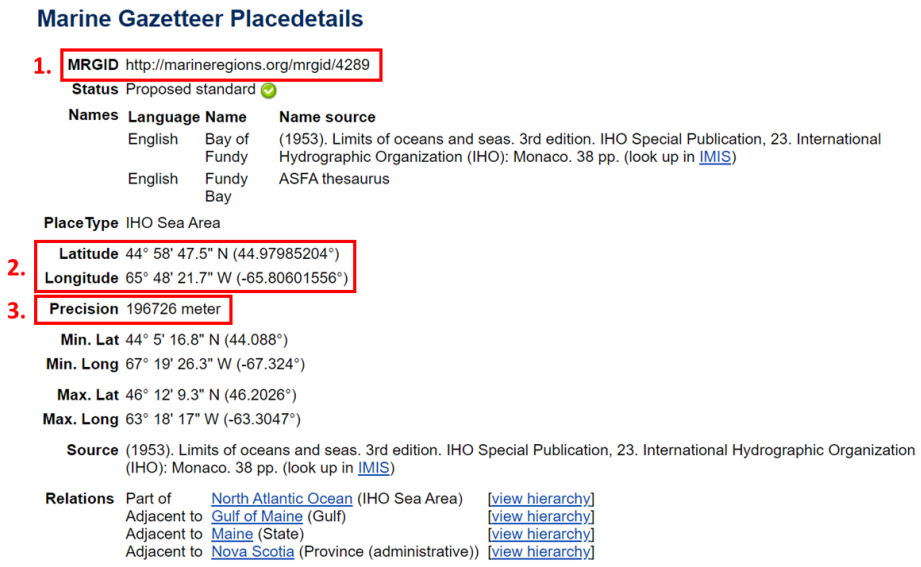


Figure 5.4: *Screenshot of Marine Region placedetails, highlighting important information for OBIS*

location portion of our dataset would look something like this:

locality	locationID	decimalLatitude	decimalLongitude	coordinateUncertaintyInMeters	georeferenceAccordingTo	georeferenceRemarks
Bay of Fundy	http://marineregions.org/mrgid/4289	44.97985204	-65.80601556	196726	Marine Regions	Coordinates are a midpoint inferred from location name

The OBIS Mapper has built-in access to the Marine Regions Gazetteer. The video below demonstrates how to use this built-in tool, as well as how to navigate the Marine Regions Gazetter to obtain important georeferencing information to include in your data.

5.6.5.1 DwC Terms obtained from Maptool and Gazetteers

Below is a table summarizing the different DwC terms you can obtain from the OBIS Maptool or from the Gazeteers discussed above.

DarwinCore Term	Maptool Term	Marine Regions Term	Notes
decimalLatitude	Latitude	Latitude	
decimalLongitude	Longitude	Longitude	
locationID		MRGID	
coordinateUncertaintyInMeters	radius	precision (not always available)	
footprintWKT	WKT		

5.6.6 Low confidence taxonomic identification

In case of low confidence taxonomic identifications, and/or the scientific name contains qualifiers such as cf., ?, or aff., then you should:

- Put the name of the lowest possible taxon rank that can be determined with high-confidence in `scientificName` (e.g. the genus)
- Put any text regarding identification with low confidence and/or qualifiers in `identificationQualifier` (e.g., cf., aff.)
- Put the species name in `specificEpithet`
- Place the rank of the taxon documented in scientificName (e.g., genus) in `taxonRank`

- Document any relevant comments in `taxonRemarks` or `identificationRemarks` (e.g. reasoning for identification)

Take an example specimen named Pterois cf. volitans. The associated occurrence record would have the following taxonomic information:

- `scientificName` = Pterois
- `identificationQualifier` = cf. volitans
- `specificEpithet` = *leave blank*
- `scientificNameID` = the one for Pterois
- `taxonRank` = genus

If the provided name is unaccepted in WoRMS, it is okay to use the unaccepted name in this field. `scientificNameID` should contain the WoRMS LSID for the genus.

There is a new Darwin Core term `verbatimIdentification` meant for containing the originally documented name, however this term is not yet implemented in OBIS so if you populate this field it will not be indexed alongside your data. However you can use `identificationRemarks` to add extra information.

The use and definitions for additional Open Nomenclature (ON) signs (`identificationQualifier`) can be found in Open Nomenclature in the biodiversity era, which provides examples for using the main Open Nomenclature qualifiers associated with physical specimens (Figure 1). Whereas the publication Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications provides examples and definitions for identificationQualifiers for image-based non-physical specimens (Figure 2).

If the occurrence is instead unknown or new to science, it should be documented according to recommendations by Horton et al. 2021. Populate the `scientificName` field with the genus, and in `identificationQualifer` provide the ON sign 'sp.'. Be sure to also indicate the reason why species-level identification is unavailable by supplementing 'sp.' with either stet. (stetit) or indet. (indeterminabilis). If neither of these are applicable, (e.g. for undescribed new species), add a unique taxon identifier code after 'sp.' to `identificationQualifer`. For example Eurythenes sp. DISCOLL.PAP.JC165.674. When adding a taxon identifier codes, pPlease avoid simple alphanumeric codes (i.e. Eurythenes sp. 1, Eurythenes sp. A). Like creating `eventIDs` or `occurrenceIDs`, try to provide more complex and globally unique identifiers. Identifiers could be constructed by combining higher taxonomic information with information related to a collection, institution, museum or collection code, sample number or museum accession number, expedition, dive number, or timestamp. This ensures namestrings will remain unique within OBIS. We also recommend including these temporary names on specimen labels for physical specimens.

5.6.6.1 Changes in taxonomic classification

Taxonomic classification can change over time - so what does that mean for your datasets when records change classification?

Because OBIS relies on WoRMS as the taxonomic backbone, changes in taxonomic classification will be updated between a day to a few weeks from the date of change, unless triggered manually. This means that the WoRMS LSID associated with a species in question will be used to automatically populate the taxonomic classification with the updated information.

However we recognize that there may be issues when larger changes occur (e.g. Family level splits). Then, records that are only identified to the Family level may not get updated properly. For example, there was a shift in coral taxonomy where the family Nephtheidae was split into Capnellidae and Alyconiidae for species occurring in the Northern Atlantic. While species that were identified as belonging to Nephtheidae have now been updated to belong to one of those two families, records that were only identified down to family (i.e., Nephtheidae) are still documented as Nephtheidae. Unfortunately there is currently no solution for dealing with splits like this, aside from contacting data providers and asking them to change the taxonomy.

A future solution may be an annotation registry that includes statements like "identifications of taxon x in geographic area y are misidentified and should be linked to taxon z".

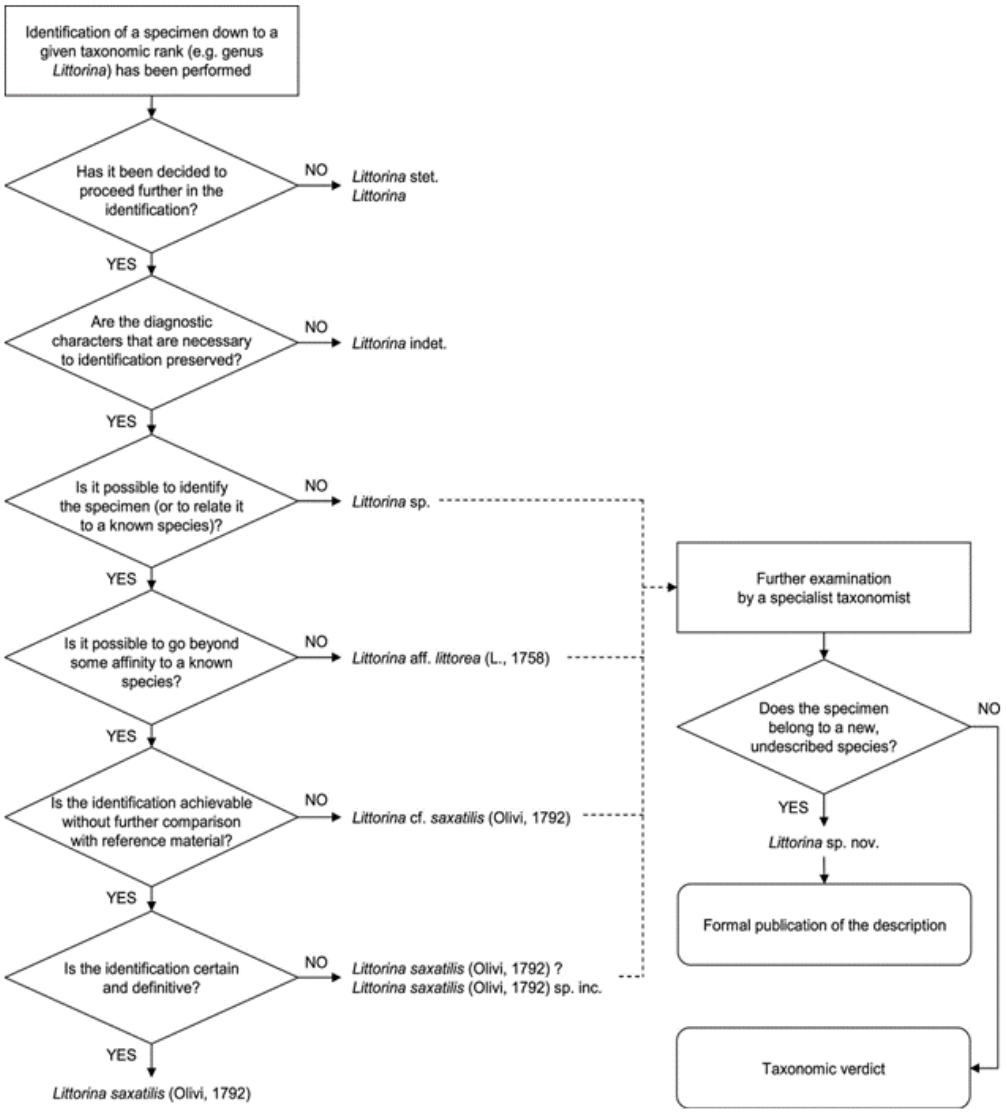


Figure 5.5: *Figure 1. Flow diagram with the main Open Nomenclature qualifiers associated with physical specimens. The degree of confidence in the correct identifier increases from the top down. More info and figure copied from Open Nomenclature in the biodiversity era.*

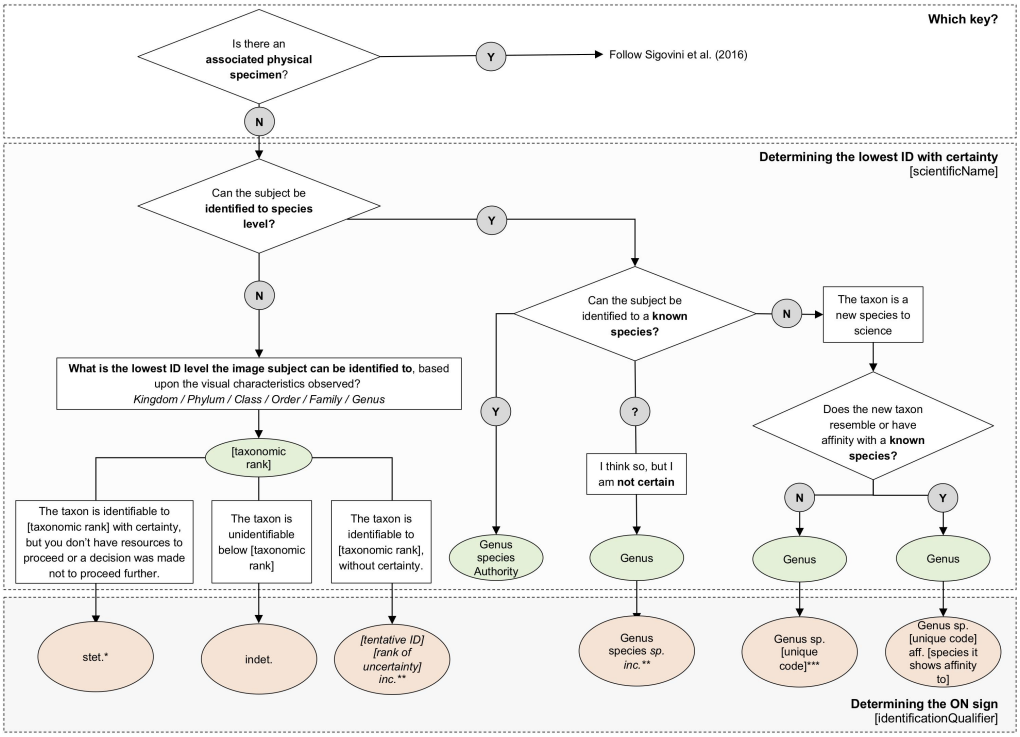


Figure 5.6: *Figure 2: Flow diagram with the main Open Nomenclature qualifiers for the identification of specimens from images (non-physical, image-based) . More information and figure copied from Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications*

5.6.7 Uncertain measurements

5.6.7.1 individualCount

In some cases `individualCount` may be uncertain due to only fragments of organisms being found, or only a range of individuals is known. The `occurrenceRemarks` can be used to document information about decisions made during data formatting.

In cases where only certain body parts (e.g., head, tail, arm) are recorded, you can incorporate this information in `measurementTypeID` by finding a P01 code that contains elements from the S12 collection, which documents sub-components of biological entities. If a P01 code for your sub-component does not exist, you can request the creation of a P01 code. If you are uncertain whether fragments are from the same individual, a range can be provided and the uncertainty can be recorded in `occurrenceRemarks`.

When providing a range to estimate the count, there are two suggested approaches:

- Use MOF/eMoF: place the data in the (extended)MeasurementOrFact extension. This requires an additional two rows per Occurrence. One with the equivalent of a minimumIndividualCount and one with the equivalent of maximumIndividualCount.
- Place data into `dynamicProperties`: Include the information in the Occurrence record itself, with no extension, and instead document it in `dynamicProperties`, with a value such as {“minimumIndividualCount”:0, “maximumIndividualCount”:5}
  - We note that documenting the data in ‘dynamicProperties’ means the information will not be machine readable and may be more difficult for users to extract.

Abundance vs Count data: A brief clarification on abundance and count data: abundance is the number of individuals within an area or volume. This type of data is recorded in `organismQuantity`, where `organismQuantityType` should be used to specify the type of quantity (e.g., individuals, percent biomass, Braun Blanquet Scale, etc.).

However, if the value is just the number of individuals without reference to a space, this information is recorded in `individualCount`.

5.6.8 Non-marine species

If you are given an error that your taxon is not marine, please confirm first whether the species is actually freshwater by cross referencing with WoRMS or IRMNG. If the species is not marine (i.e. belongs to a non-marine genus), check with the data provider as necessary for possible misidentification. Finally you can contact WoRMS at info@marinespecies.org to discuss adding the taxon to the WoRMS register. You will be required to provide documentation in this case to confirm marine status of the taxon.

Otherwise, records marked as non-marine will be dropped from the published dataset, and this will be flagged in the data quality associated with your dataset.

Let’s consider an example within this dataset on benthic macroalgae. Inspecting the data quality report we can see there are three dropped records due to species not being marine.

Clicking on the dropped records we can see which three species were dropped. By scrolling to the right of the table, we can see these records have two quality flags: NO_DEPTH and NOT_MARINE.

Let’s take a look at the first species, Pseudochantransia venezuelensis. When we search for this species on WoRMS we can see that the species is marked as freshwater.

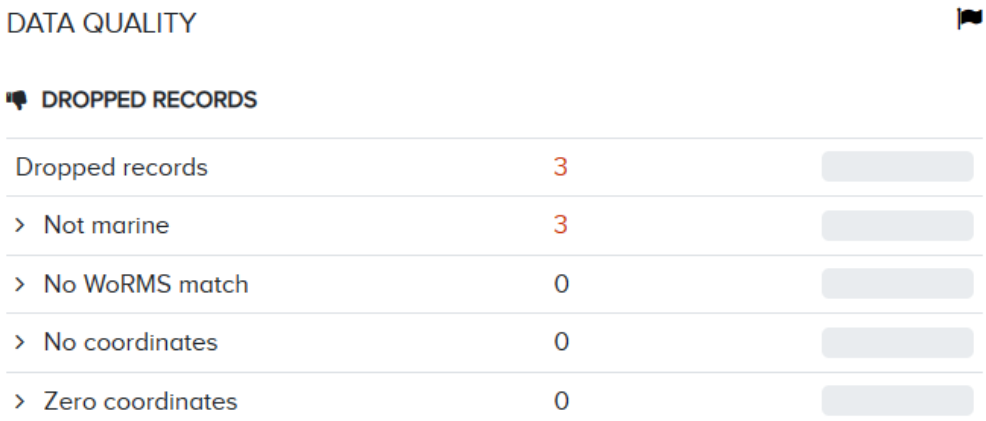


Figure 5.7: *Dropped records from a benthic macroalgae dataset*

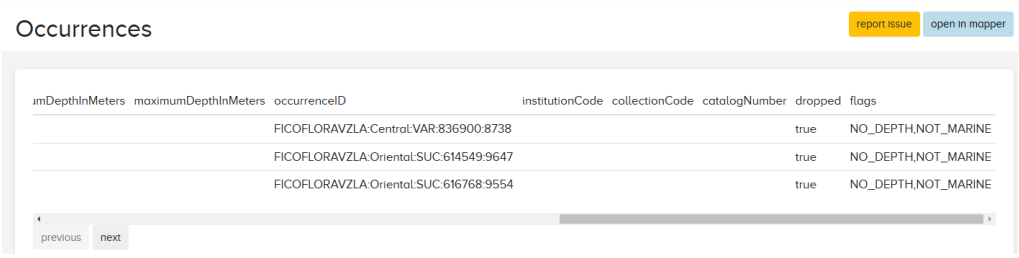
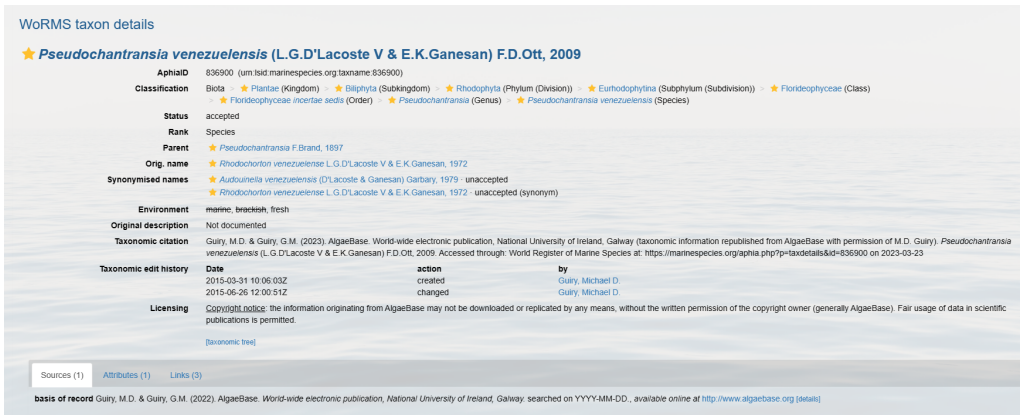
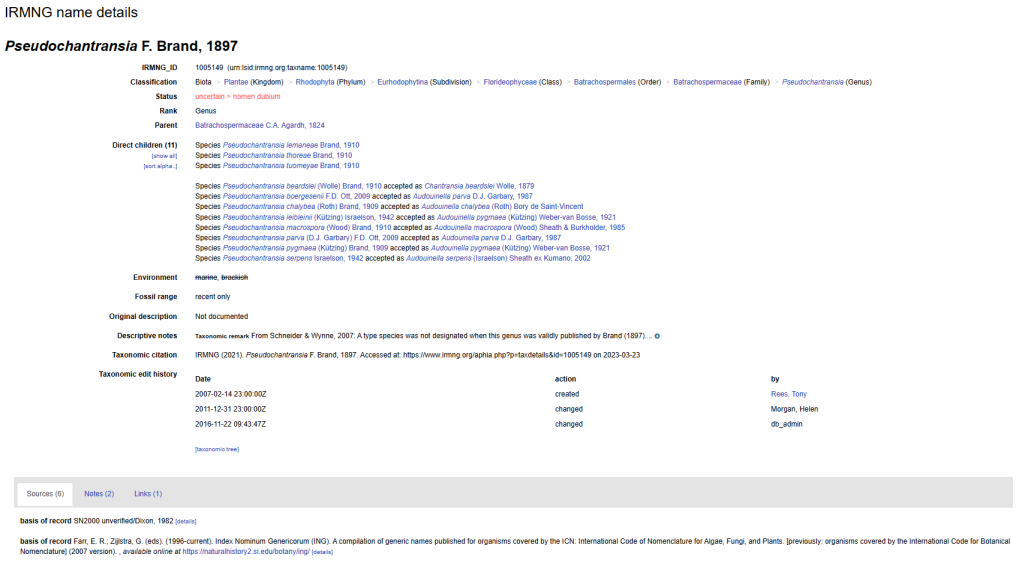


Figure 5.8: *Flags specifying why certain records were dropped*



Cross-referencing with IRMNG, if we search for the genus-species, the species is not even found, an indication that it is not in the database (and also why it can be good to check multiple sources). Searching for just the genus, we can see that marine and brackish are stricken out, indicating the species is not marine.



If you have species that are marked as non-marine in these registers but are either supposed to be marine, or were found in a marine environment, then you should contact WoRMS to discuss adding it to the register. For additions and/or edits to environmental or distribution records of a species, contact the WoRMS Data Management Team at info@marinespecies.org with your request along with your record or publication substantiating the addition/change.

Publishing Data

Chapter 6

Data publication and sharing

Once you have finished formatting your data and have conducted some basic quality control procedures on it, you are ready to publish your dataset to OBIS. Note that OBIS nodes can accept any data files from its data sources or data providers, and they publish these data on their respective OBIS node Integrated Publishing Toolkit (IPT). Data from node IPTs are then harvested by central OBIS to become accessible via the global database. The Integrated Publishing Toolkit (IPT) is developed and maintained by the Global Biodiversity Information Facility (GBIF). While GBIF maintains an IPT manual, we outline specific OBIS instructions here.

Some nodes require the **manager themself to upload data**. This means after you have formatted your data for OBIS and have done some basic quality control steps, you simply pass this file on to the appropriate node manager. They will communicate with you any issues during the publication process.

Sometimes a node manager will request the **data provider to upload their own dataset** to the IPT. If you the data provider are required to upload your data then there are a number of steps you will take to upload your data. A reminder that all data formatting and quality control steps should be completed before uploading your dataset. Once uploaded, the node manager will check your upload before publishing the dataset.

There are a few steps involved to publish a dataset, whether you are the provider or the node manager. You must:

1. Identify the correct IPT for your OBIS node (you may have to contact your node manager to confirm IPT)
2. Login to the IPT, or have the node manager create an account for you if you do not have one, so you can upload your dataset(s)
3. Map each of your fields to Darwin Core terms. This should be relatively straight forward if you have done this already during data formatting
4. Fill all relevant metadata to help users understand and cite your dataset
5. Publish and make your data public

Details for each of these steps are outlined on the subsequent pages. We will start with an overview of which Creative Commons license should be used, and then move on to using the IPT to publish data.

6.1 Licenses

OBIS nodes must make the necessary agreements with the original data providers so that data can be made available to OBIS under one of the following Creative Commons licenses (in order of preference):

- CC-0 - data may be used without restrictions
- CC-BY - data are available for any use if proper attribution and credit is given
- CC-BY-NC - data may be used for any non-commercial use as long as proper attribution/credit is given

You may need to consult with your organization if there are any copyright concerns. For more information on the different Creative commons license types see About the licenses.

6.2 IPT: Integrated Publishing Toolkit

Contents:

6.2.1 Introduction to the IPT

Before we get into the details for accessing and using the IPT, let's understand what it is. Biodiversity datasets and their metadata are published in OBIS using the Integrated Publishing Toolkit (IPT), developed by GBIF. The IPT is an open source web application that can be customized by the OBIS node manager (see IPT admin page for details). An IPT-instance is used to publish and register all datasets. To be able to create and manage your own dataset (called a "resource" by GBIF), you will need a user account. In general, the IPT software assists users in mapping data to valid Darwin Core terms, as well as archiving and compressing the Darwin Core content with:

1. A descriptor file: `meta.xml` that maps the core and extensions files to Darwin Core terms, and describes how the core and extensions files are linked
2. The `eml.xml` file, which contains the dataset metadata in Ecological Metadata Language (EML) format. For instructions on how to enter the metadata go to EML.

All these components (i.e., core file, extension files, descriptor file, and metadata file) become compressed together (as a .zip file) and comprise the Darwin Core Archive.

6.2.2 How to access the IPT

Once you have determined which OBIS node IPT is suited for your dataset, you can contact your node manager to create an associated account for you. There will be a link on the sign in page that will direct you to the IPT's administrator to contact them. If your node's IPT is not listed here, you will have to contact the node manager to get the link to their IPT.

If you are an IPT admin and want to know how to set up an IPT yourself, see the IPT admin page.

6.2.2.1 Who populates the IPT with datasets?

With regard to populating the IPT with marine data for OBIS, there are two possible approaches:

1. Manager driven: You as node manager take the responsibility of describing, checking and uploading the data and metadata to the IPT. The data provider can send you the data 'as such' or you can make agreements with your providers on the accepted OBIS data format and standards. This approach will

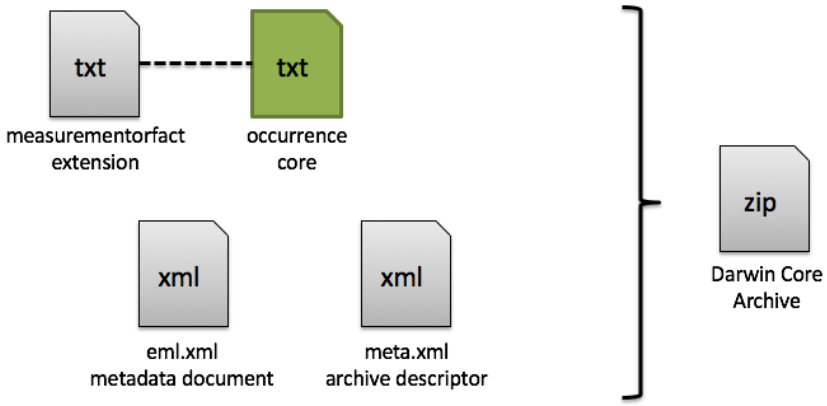


Figure 6.1: *Example showing how Occurrence core, EML, and meta.xml files make up a Darwin Core-Archive file*

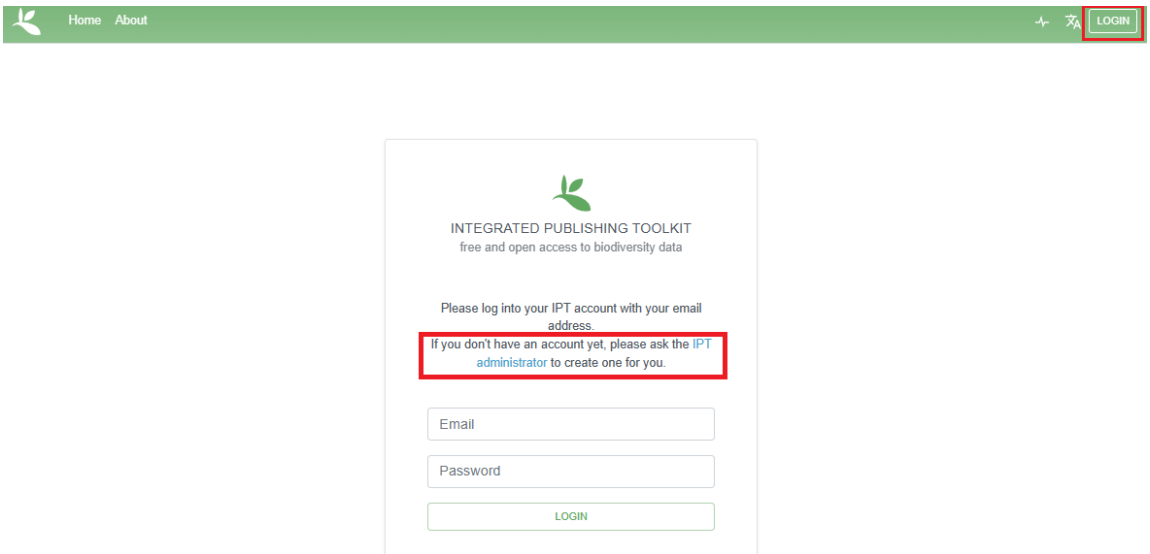


Figure 6.2: *Screenshot of IPT login page, highlighting link to IPT admin and the login button*

give you a very good knowledge of what data is available. It can be time-consuming, as (extended) communication with the data provider will be necessary to document the metadata and to re-format the data to the OBIS standards.

2. User driven: You as node manager can guide (some of your) data providers to publish the data and metadata to the IPT themselves. Your main task will be to make sure that all relevant information and data for OBIS is available and that you perform the necessary quality checks before the data are released to OBIS. Once the Darwin Core Archive is created, the data provider should inform the node manager of this action, so he or she can do the necessary quality control checks. In order for the node manager to be able to look at the dataset, the data provider should add him or her as a "resource manager" to this specific dataset.

In most cases, there will be a combination of these two approaches. The chosen approach will largely depend on the capacity, availability and willingness of your data provider to invest extra time in formatting and thoroughly describing their data. If you – as node manager – would prefer a partly user driven approach, the following steps to publishing marine data to OBIS briefly explains how you or a data provider can upload, standardize and publish a dataset on the OBIS node IPT, without the hassle of installing and maintaining an IPT instance. The data are published in your organization's name. This guide is based on the Canadensys 7-step guide to publishing marine data:

Desmet, P. & C. Sinou. 2012. 7-step guide to data publication. Canadensys. http://community.canadensys.net/publication/data-publication-guide.

Caution: Make sure you have obtained the rights from the data owners to publish their data!

6.2.3 Create your resource on the IPT

Once you have your account, login at the top of the IPT page. Click on the tab Manage resources: it will display all the datasets you are managing and will be empty at first. You can create a new resource at the bottom of the page. Follow the GBIF IPT manual for more detailed instructions.

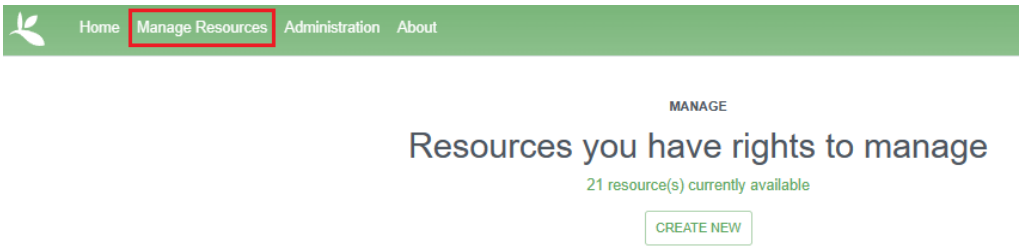


Figure 6.3: *Screenshot of the manage resources page on the IPT*

The first thing that needs to be completed is the shortname of your resource. This shortname uniquely identifies your resource (i.e. dataset) and will eventually show up in the URL of this resource on IPT. These shortname identifiers are also used to create folders on the IPT and **they cannot be changed**.

We therefore advise that the shortname:

- is unique, descriptive and short (max. 100 characters)
- does not contain a space, comma, accents or special characters

Shortname good examples:

- VLIZ_benthos_NorthSea_2000
- UBC_algae_specimens
- ...

When you would delete a resource, please inform your node manager of this action! If you create a test-file, please include `_test` at the end of your shortname.

Then select the type of data you are uploading: Occurrence, Checklist, Sampling-Event (i.e. Event Core), Metadata only, or Other. Note that Checklist datasets are accepted by GBIF, but not currently implemented in OBIS. However, you can still have checklist data hosted on OBIS IPTs.

You can also create an entirely new resource by uploading an existing archived resource. See the IPT manual section Upload a Darwin Core-Archive for instructions.

Please note the IPT has a 100MB file upload limit, however, there is no limit to the size of a Darwin Core Archive that the IPT can export/publish. Refer to this note in the IPT manual to find out how to work around the file upload limit.

Once you have created your resource, you will see an empty resource overview page.

6.2.3.1 Upload data

Uploading your source file to the IPT is easy: go to > your resource overview page > Source Data and click on Choose File. This is where you will select and add the files containing your Core table and (if applicable) extensions.

You might want to compress/zip your source file first to improve the upload speed of large files. The IPT will unzip them automatically once received. Follow the IPT manual for more detailed instructions (including the option to use multiple source files or to upload via a direct database connection).

Accepted formats are delimited text files (csv, tab and files using any other delimiter), either directly or compressed as zip or gzip.

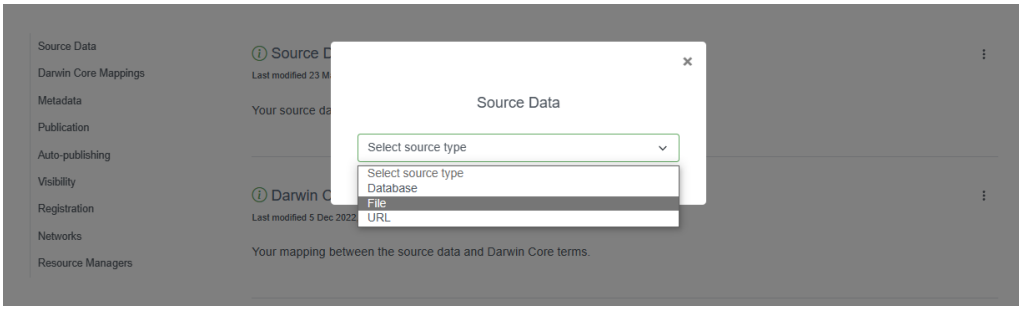


Figure 6.4: *Screenshot of where on the IPT page you can add source data*

Let's follow an example Event core dataset. The first file we will upload will be a .csv file containing the Event core table. After it's been selected, we click "Add".

A source file detail page will then be shown, displaying how the IPT has interpreted your file (number of columns, rows, header rows, character encoding, delimiters, etc.). Click the preview button to verify everything is correct, click anywhere on the screen to exit the preview, then click save.

You can also provide information about how the data table is encoded, how many header rows exist, the type of delimiters, and what type of character encoding you used. You may need to double check the character encoding for your file, especially if you used any special characters (e.g., in species or place names). **UTF-8** is one of the most common encoding standards, and you can select this encoding when saving files, depending on the software used:

- Windows: MS Excel: select Save as. From the drop down select the "CSV UTF-8 (Comma delimited)" option
- Windows: Notepad: Click on File, then Save as

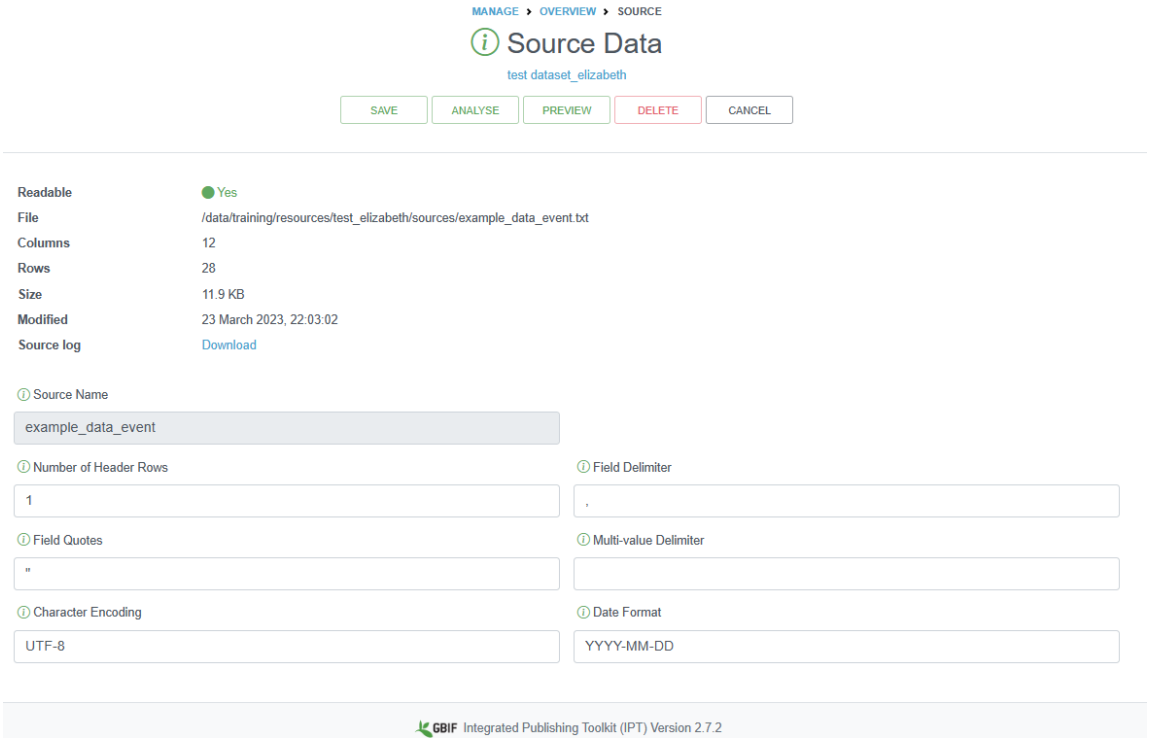


Figure 6.5: *Screenshot of page shown after you upload a source file for the first time*

  - In the drop down menu "Save as type" drop-down, select All Files
  - In the "Encoding" drop-down on the bottom right, select UTF-8.
  - Be sure to name your file using the .csv extension (e.g., data.csv).
- Mac: Numbers: Select File then Export to... -> CSV.
  - Click Advanced options
  - Click the drop-down menu next to Text Encoding and select Unicode (UTF-8)
  - Click Next, then finally name your file, select a location, and click Save

Once you have specified all the above information, click Save, This will redirect you back to the Overview page for your dataset. You can add multiple files for each of your tables by clicking the three vertical dots to the right and "+ Add". We will add a .csv file for an Occurrence extension, as well as one for the eMoF extension.

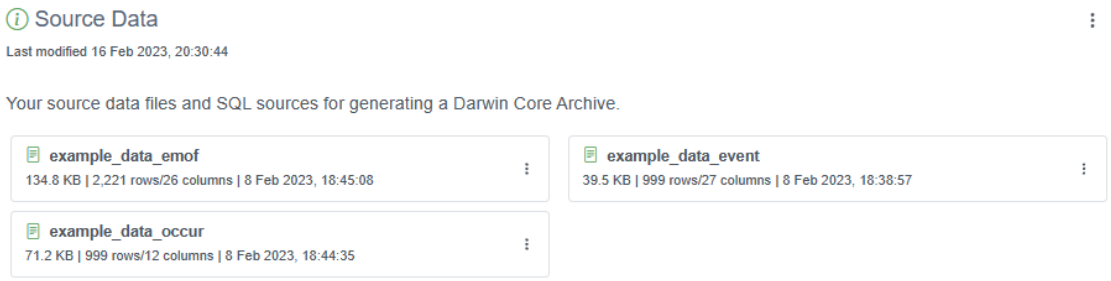


Figure 6.6: *Example screenshot of the different files, core and extensions, you can upload to an IPT*

6.2.4 Map your data to Darwin Core

Because biodiversity data are published in the Darwin Core standard, the next step is to map your data fields to Darwin Core. As we have mentioned earlier in this manual, the DwC standard includes a list of defined terms and allows your data to be understood and used by others. It also allows an aggregator like OBIS or GBIF to integrate your data with other datasets.

Darwin Core mapping is the process of linking the fields in your resource file with the appropriate Darwin Core terms. It is the most challenging step in publishing your data for two reasons:

1. The list of Darwin Core terms can be overwhelming, so it might be difficult to select the ones that are appropriate for your dataset
2. The IPT currently only allows one-to-one mapping of fields, so the ease of mapping will depend on your database structure and on the feasibility of exporting as close to Darwin Core as possible. You can contact your node manager or the OBIS secretariat at info@iobis.org to help guide you through the steps, review your mapping, suggest terms etc. You also welcome to post questions in the OBIS Slack.

You can find more information regarding Darwin Core mapping in the IPT manual (including core types, extensions, auto-mapping, default values, value translation, etc.).

To add the DwC mappings, click the three vertical dots to the right of this section and select "+ Add"

Figure 6.7: *Screenshot of the IPT demonstrating where to add new DwC mappings*

A popup window will allow you to select your Core type to facilitate mapping - Occurrence or Event core. In this example, we will select Darwin Core Event for the Event core table.

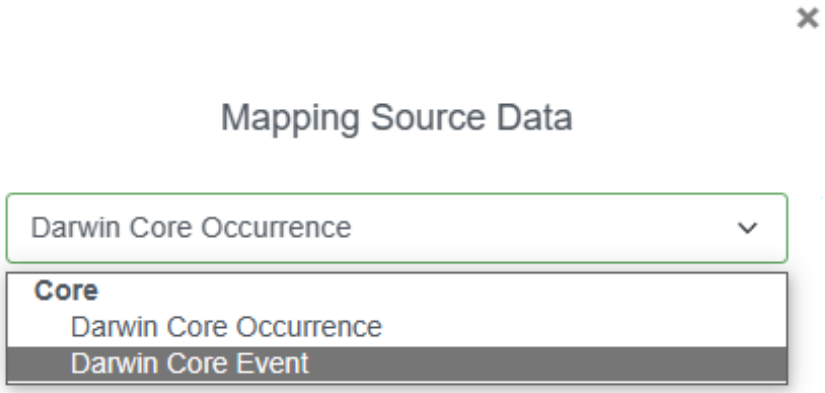


Figure 6.8: *Screenshot showing the two Core types you can map a datafile to*

Next we will confirm which file we are mapping to: the .csv file containing our event data. Then click Save.

Because we had already named several of our fields with Darwin Core terms, they have been auto-mapped for us. It is important to review each Darwin Core class (Record, Event, Location, etc.) and ensure any unmapped fields are mapped to the correct DwC term. When you select a term, a drop down menu will appear where you

can select the appropriate field from your dataset. It is good practice to double check that the auto-mapped fields are mapped correctly.

Once you are done mapping, any unmapped fields will appear at the bottom of the page for you to check. If there is no DwC term to map these terms to, that is okay, but the data will not be published alongside the rest of your dataset. Consider moving these unmapped fields to either `dynamicProperties` or to one of the extensions (e.g., eMoF), whichever is most applicable.

Finally, click Save. You may return to the Overview page by clicking Back. To add DwC mappings for the other files (Occurrence and eMoF), click the same "+ Add" button and go through the same process for each extension table you have.

The IPT may identify Redundant terms if certain terms appear in the e.g., Event core and Occurrence extension. If your Occurrence extension (or core) contains information about individualCount and organismQuantity, you can map such fields in both the Occurrence and the eMoF as a measurementType.

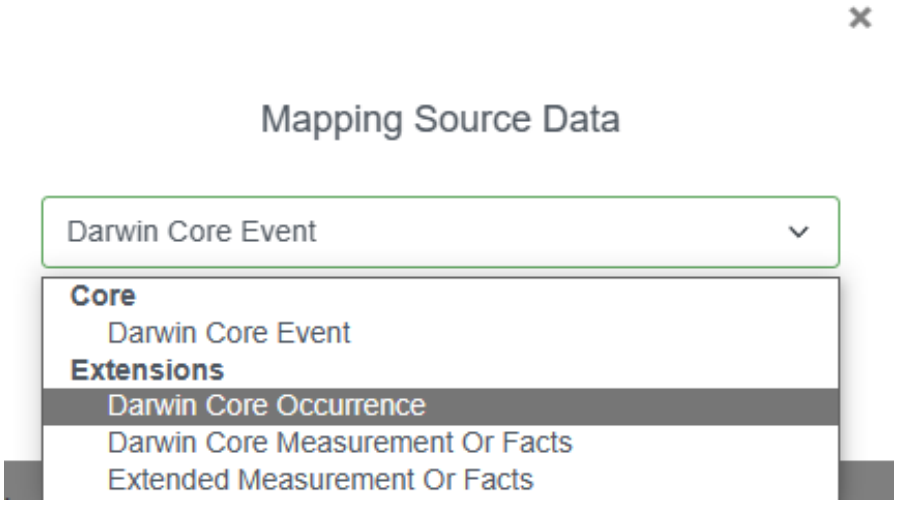


Figure 6.9: *Screenshot showing other DwC extensions you can map to*

Watch the video below for an overview of all of the above procedures, from uploading data to mapping terms to DwC.

The next step is to fill in or upload metadata.

6.2.5 Add metadata

Metadata enables users to discover, assess, understand and attribute your dataset for their particular needs, so it pays off to invest some time providing them.

Go to your resource overview page > Metadata and click Edit to open the metadata editor. Any information you provide here will be visible on the resource homepage and bundled together with your data when you publish.

Follow the guidelines on the OBIS metadata standards and best practices page, or check the IPT manual for detailed instructions about the metadata editor. You can also upload a file with metadata information. The video below also demonstrates how to fill metadata on the IPT.

Figure 6.10: *Screenshot showing where to add or upload metadata*

6.2.6 Publish on the IPT

With your dataset uploaded, properly mapped to DwC, and all the metadata filled, you can publish your dataset. On your resource overview page, go to the Publication section, click the vertical dots and select Publish.

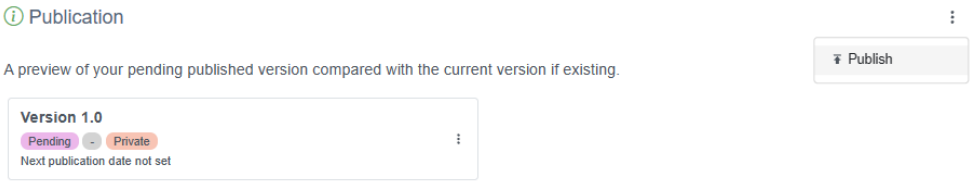


Figure 6.11: *Screenshot showing where to manage the publishing of your dataset*

The IPT will now generate your data as Darwin Core, and combine the data with the metadata to package it as a standardized zip-file called a "Darwin Core Archive". See the IPT manual for more details.

> **Note:** Hitting the "publish" button does not mean that your dataset is available to everyone, it is still private, with access limited to the resource managers. It will only be publicly available when you have changed Visibility to Public. You can choose to do this immediately or at a set date.

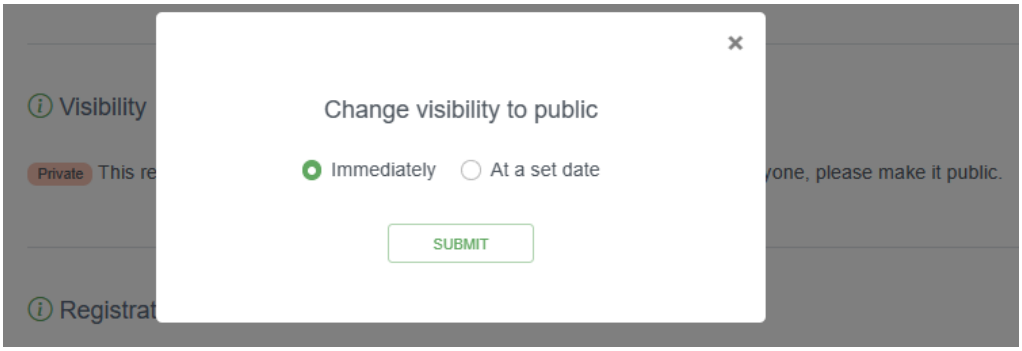


Figure 6.12: *Screenshot showing how to change the visibility of your dataset*

Your dataset will only be harvested by GBIF when you change Registration to Registered. This step is not needed for OBIS to harvest your datasets. Please do not register your dataset with GBIF if your dataset is already published in GBIF by another publisher. Note that the IPT itself must be registered with GBIF in order to publish to GBIF. The node manager can do this.

Back on the resource overview page > Published Release, you can see the details of your first published dataset, including the publication date and the version number. Since your dataset is published privately, the only thing left to do is to click Visibility to Public (see the IPT manual) to make it available to everyone.

> **Warning**: please do **not** do this with your test dataset.

Figure 6.13: *Screenshot showing where to register your dataset with GBIF. This is only available if the IPT itself is registered with GBIF as well.*

It is now listed on the IPT homepage and you can share and link to it, e.g.: `http://ipt.vliz.be/resource.do?r=kielbay70`. This would be a good time to notify any regional or thematic network you are involved in, which can also have an interest in your dataset.

Your published dataset is a static snapshot of your data and will not change until you upload an updated source file and click publish again or publish a new version (do not create a new resource). This procedure has the advantage that your dataset is always available, does not require a live internet connection to your database and can be easily shared. It also allows you to control the publication process more precisely: version 1, version 2, etc. and users are informed of how recent the data are (via the last publication date).

To view an older version of the metadata about the resource, just add the trailing parameter `&v=n` to the URL where `v` stands for "version", and `n` gets replaced by the version number, e.g., `http://ipt.vliz.be/ilvo/resource.do?r=zoopl_bpns&v=1`. In this way, specific versions of a resource's EML, RTF, and DwC-A files can be retrieved. Please note, the IPT's Archival Mode must be turned on in order for old versions of DwC-A to be stored (see Configure IPT settings section of the IPT manual).

The first minute of the video below provides an overview of how to publish on the IPT.

6.2.6.1 Publish your metadata as a data paper

The Metadata expressed in the EML Profile standard can also be downloaded as a Rich Text Format (RTF) file. The latter can serve as a draft manuscript for a data paper (First database-derived 'data paper' published in journal), which can be submitted for peer-review to e.g. a Pensoft journal.

6.2.7 Downloading datasets from an IPT

To download a dataset from an IPT, simply login, and from the home page (not the Manage Resources tab) search for the dataset in question. You can search for keywords in the Filter box on the right side of the page.

Once you navigate to the page of a dataset, at the top of the page you will have options to download the whole Darwin Core Archive file, or just the metadata as an EML or RTF file.

6.3 IPT Administration Responsibilities

If you are an IPT administrator, you are responsible for:

1. Creating the IPT
2. Registering IPT with OBIS and GBIF
3. Keeping the IPT up to date
4. Describing, checking and uploading the data and metadata to the IPT
5. Guide data providers to publish the data and metadata to the IPT themselves
6. Make sure all relevant information and data for OBIS is available
7. Perform necessary quality checks before the data are released to OBIS

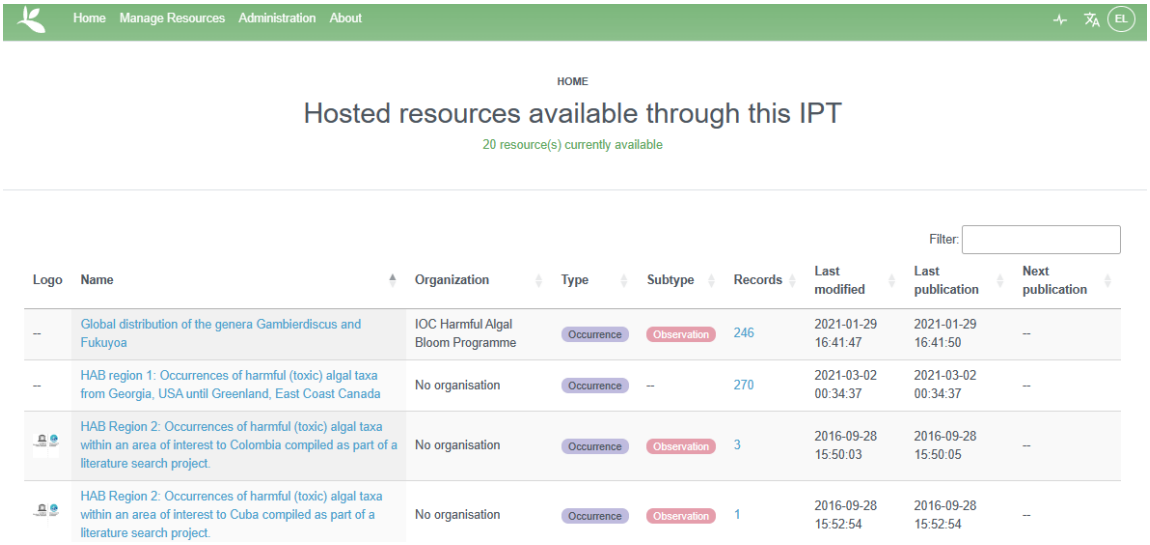


Figure 6.14: *Overview of home page of an IPT*

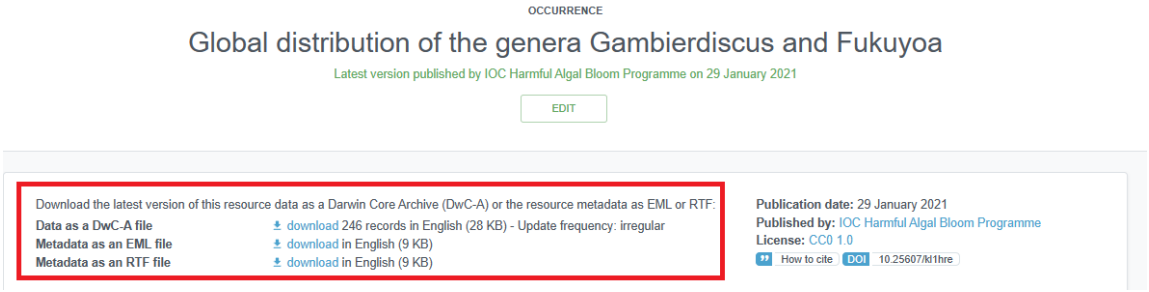


Figure 6.15: *Overview of a dataset page on an IPT, emphasizing where to download the resource*

6.3.1 Creating and Installing IPTs

OBIS nodes can decide to install and manage their IPT on their own institutional servers or use (at no charge!) the OBIS servers in Oostende, Belgium, provided as in-kind by the Flanders Marine Institute (VLIZ), which also runs the European OBIS node (EurOBIS). VLIZ also ensures the IPT instances run on the latest version (important for security updates). Here is an overview of the IPT instances hosted in Oostende: http://ipt.iobis.org/. Please contact the secretariat at info@iobis.org if you would like OBIS to host your IPT.

To install your own IPT, please follow the instructions in the GBIF IPT manual.

6.3.2 Registration

When you have installed your IPT, please provide the IPT instance URL to the OBIS secretariat (helpdesk@obis.org), so your IPT is included in the data harvesting process.

OBIS recommends to share the data as widely as possible including with other networks such as GBIF. On 13 October 2014, a cooperation agreement was signed between the secretariats of IOC-UNESCO/OBIS and GBIF in which the two parties recognized the two initiatives (OBIS and GBIF) as complementary with common goals (and in particular OBIS's role in Marine Biodiversity Data). Together they agreed to work towards maximizing the quantity, quality, completeness and fitness for use of marine biodiversity data, accessible through OBIS and GBIF and in particular in the development of data standards (DwC), technology (IPT), maximizing fitness for use, development of biodiversity indicators for assessments, enhance capacity through training and coordinate approaches to the global science/policy interface. At the 4th session of the OBIS Steering Group (SG-OBIS-IV, Feb 2015), it was recommended that GBIF should harvest OBIS tier 2 nodes if OBIS tier 2 nodes could also harvest marine datasets from their GBIF nodes. In this way OBIS could work directly with the entire marine community and promote its standards and best practices. It was not recommended that iOBIS set up a separate IPT for GBIF to harvest, since this would mean a duplication of effort.

In order to publish data with GBIF, the OBIS node also needs to become a data publisher in GBIF, and link the IPT installation with this publishing organization. OBIS nodes are encouraged to use the OBIS node name as the publishers's name, unless the host institution requires its institutional name to be used. In the latter case, reference to the OBIS node can be added in the description, as well as between brackets in the title. The name of the IPT instance can also refer to the OBIS node. OBIS nodes are also encouraged to select OBIS as the endorsing organization. In this way, the OBIS node is also listed on the OBIS page at GBIF.

The video tutorial below will help guide you on registering your OBIS node IPT with GBIF.

6.3.3 Maintaining the IPT

As the IPT administrator, you must make sure the IPT is kept up to date, datasets (resources) are published in a timely manner, add, edit, or delete IPT users as required, register with GBIF as applicable, and configure the types of cores and extensions accepted by the IPT.

To add or update extensions, navigate to the Darwin Core Types and Extensions page from the Administration menu. To install an extension (e.g., DNA Derived Data), simply scroll down the page and click the `Install` button to the right of the desired extension.

For extensions already installed, you may notice yellow flags indicating a core or extension is out of date. You can update these easily by clicking the `Update` button.

For a detailed breakdown of administrator options, see the IPT guide.

6.4 Maintaining and sharing published data

Content

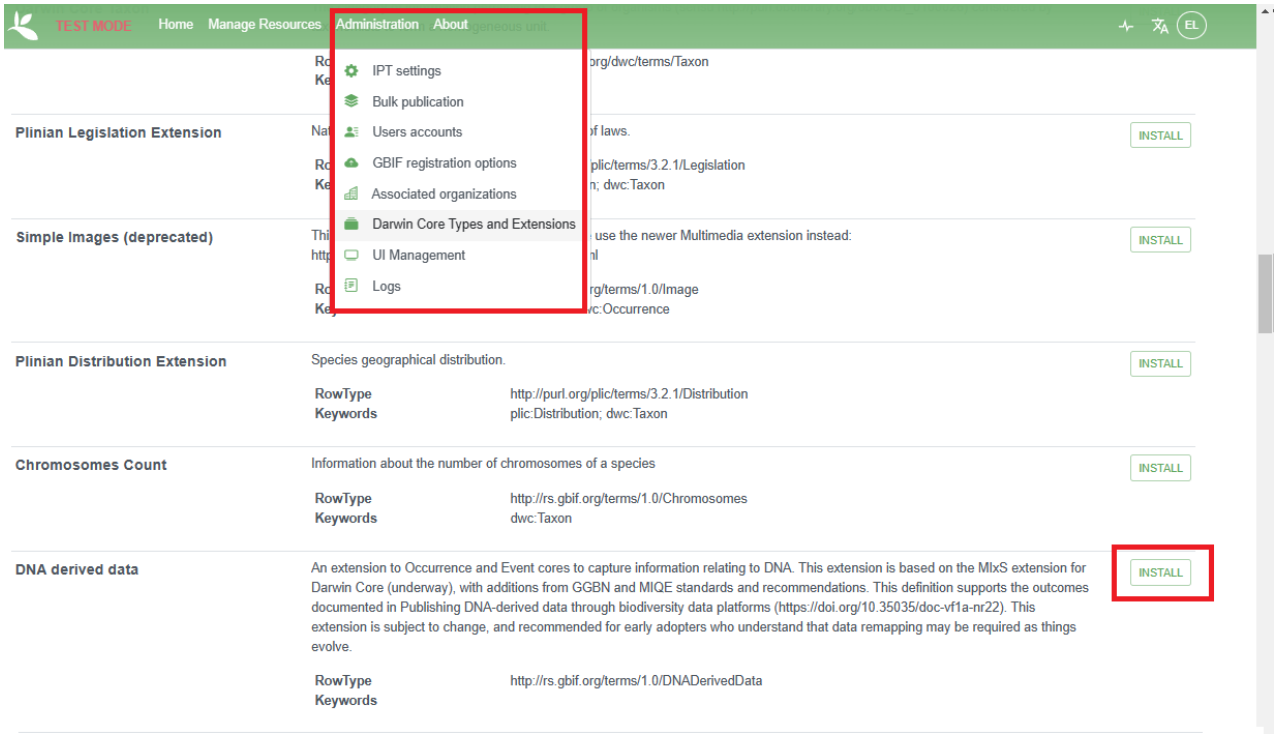


Figure 6.16: *Screenshot of IPT Admin page*

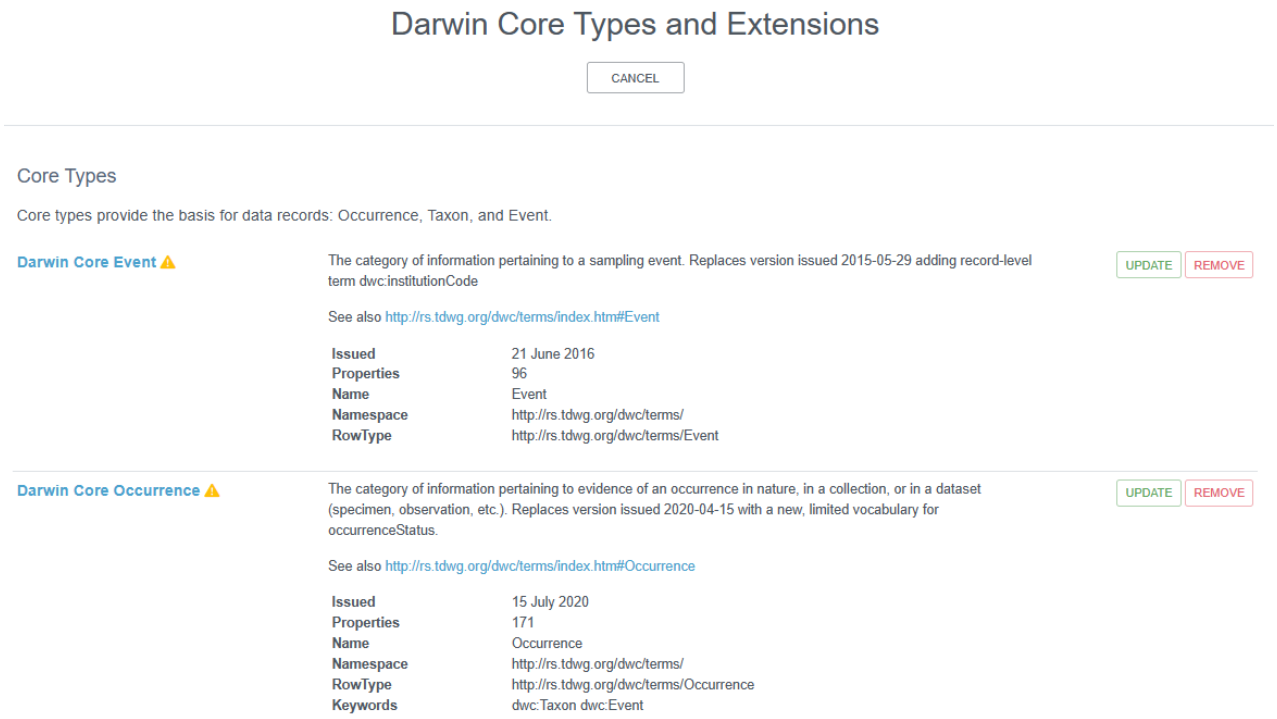


Figure 6.17: *Screenshot demonstrating when core or extensions need to be updated*

- User tracking
- Update your own data
- Publish to OBIS and GBIF

6.4.1 Adding a DOI to datasets

DOIs are important for tracking your dataset. Fortunately you can easily reserve a DOI for your dataset if the IPT administrator has configured the IPT accordingly.

As the IPT administrator, you must enable the capacity for users to reserve DOIs. To do this you first need a DataCite account associated with an Organization. Only one DataCite account can be used to register DOIs in this manner (i.e. IPT users do not need an account). The IPT's archival mode, configurable on the IPT settings page, must also be turned on (note that enabling this mode will use more disk space) to enable this feature. For more information see the IPT administration manual.

Once this has been configured, a data provider or admin can easily reserve a DOI for a dataset. First log in to the IPT, navigate to the Manage Resources tab, then select the dataset for which you wish to reserve a DOI. On the overview page for the dataset, scroll to the Publication section, click the three vertical dots and select "Reserve DOI".

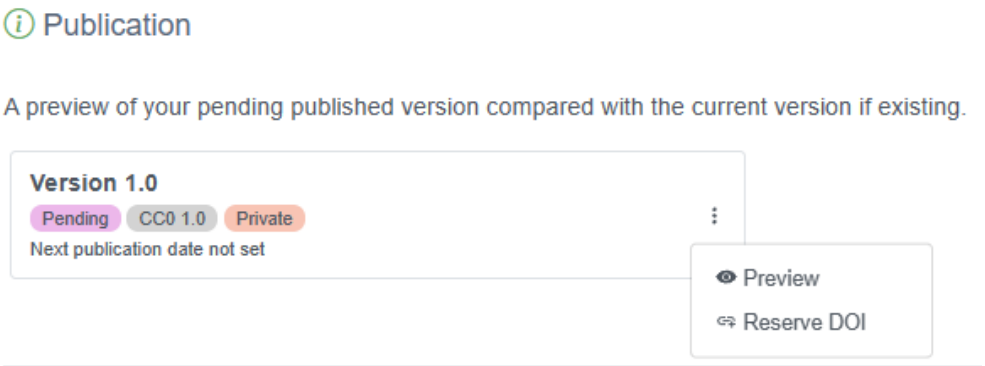


Figure 6.18: *Screenshot indicating how to reserve a DOI for your dataset*

6.4.2 User tracking

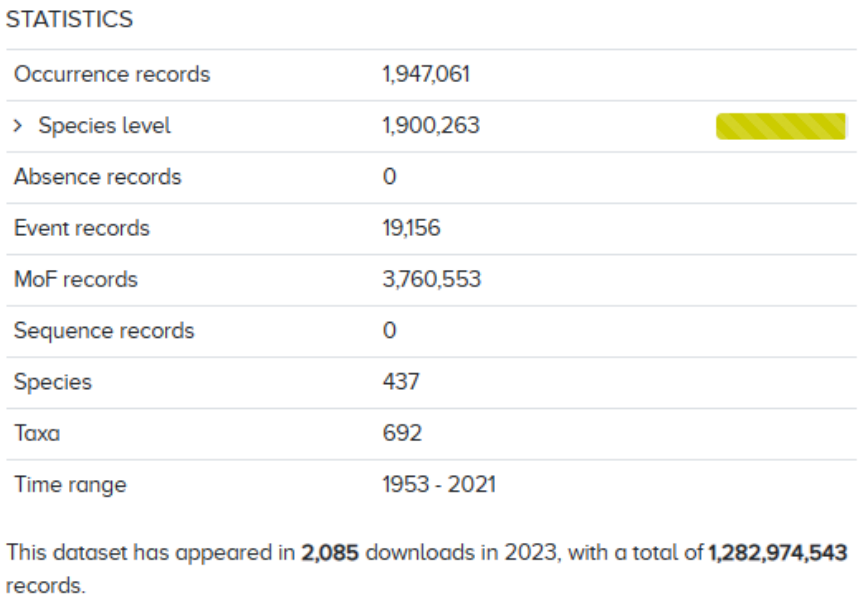
OBIS tracks the number of times your dataset is downloaded. This information is available on your dataset's page under the Statistics box.

6.5 Update your data in OBIS

To update your own data in OBIS, the process is largely the same as publishing your first version. Follow the steps below:

- Log in to the IPT where your data are hosted
- Under the Manage Resources tab, locate your dataset
- Upload new files, complete the DwC mapping, and/or update any metadata that may have changed
- In the Publication section, click Publish just as you did before

The new version will be automatically updated. A new DOI will be generated only if you generate it yourself on the new version. Deciding when to generate a new DOI is up to you, but generally you should generate a new DOI when there have been major changes to your dataset, such as significant changes in your metadata or a move to a 2.0 resource version within the IPT.



STATISTICS

Occurrence records	1,947,061
› Species level	1,900,263
Absence records	0
Event records	19,156
MoF records	3,760,553
Sequence records	0
Species	437
Taxa	692
Time range	1953 - 2021

This dataset has appeared in **2,085** downloads in 2023, with a total of **1,282,974,543** records.

Figure 6.19: *Example screenshot of how dataset downloads can be tracked*

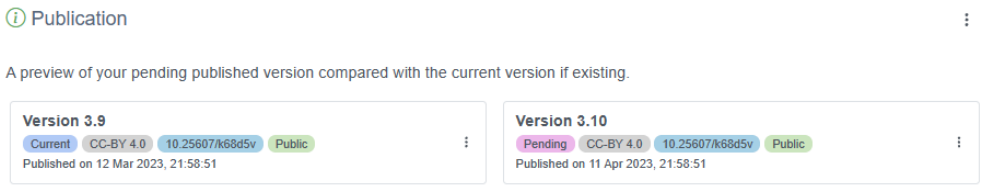


Figure 6.20: *Example of IPT version control*

6.6 Simultaneous publishing to GBIF

There are a few differences between OBIS and GBIF. Perhaps the most obvious difference is that OBIS focuses on marine data, whereas GBIF includes broader biodiversity data. However, OBIS implements more strict quality control requirements for published datasets. A complete list of GBIF's quality control checks can be found here, and a guide on GBIF's publishing process is here.

OBIS currently accepts two core data table types: Occurrence core and Event core. GBIF includes both Occurrence and Event core, with one additional core type, Checklists. GBIF is also developing a new data model to expand data publishing capabilities.

Some of the other main differences in how OBIS and GBIF structure and publish datasets that you should be aware of include:

- OBIS uses WoRMS as the exclusive taxonomic backbone (and WoRMS identifiers to populate `scientificNameID`), whereas GBIF uses Catalog of Life and does not currently require the use of taxonomic identifiers
- The OBIS-ENV-DATA structure, the eMoF extension, and the DNA Derived data extensions are not included in GBIF downloads (e.g., this dataset description). This data can still be published alongside your dataset, and is available when it is downloaded from the Source archive, but it will not be included in a GBIF Annotated Archive download.
- OBIS conducts some QC procedures that GBIF does not, including:
  - Checking validity of depth measurements
  - Checking validity of WoRMS LSID
  - Identifying if taxa are exclusively freshwater or terrestrial
- GBIF includes most of the same data standards as OBIS (Darwin Core, EML), however GBIF also follows the Biological Collection Access Service (BioCASE/ABCD)

Watch the video below for details on how to publish OBIS datasets to GBIF (starting at 1:07), or how to publish GBIF datasets to OBIS (starting at 6:42).

See the tables below for a quick comparative reference on which terms are required or recommended in OBIS and GBIF for Occurrence and Event tables.

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

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

Term	Status in OBIS	Status in GBIF
informationWithheld	not required, accepted	not required, accepted
dataGeneralizations	not required, accepted	not required, accepted
country	not required, accepted	not required, accepted

Access Data from OBIS

Chapter 7

Data access

OBIS has over 100 million records of marine data accessible for downloading. To download data from OBIS, there are several options:

- OBIS homepage or advanced dataset search
- OBIS Mapper
- Accessible through the R package robis
- OBIS API
- Full data exports
- IPT

NOTE When you download data from the Mapper or full export, the data you will receive is flattened into one table with occurrence plus event data. eMoF data tables are separate upon request. However when you download a dataset from the OBIS homepage or dataset page, all tables (Event, Occurrence, eMoF) are separate files.

7.1 OBIS Homepage and dataset pages

From the OBIS homepage, you can search for data in the search bar in the middle of the page. You can search by particular taxonomic groups, common names, dataset names, OBIS nodes, institute name, areas (e.g., Exclusive Economic Zone (EEZ)), or by the data provider's country.

When you search by dataset you will notice an additional option appears for advanced search options. This will allow you to identify specific datasets, and apply filters for OBIS nodes and whether datasets include extensions.

Regardless if you found a dataset through the homepage or the advanced Dataset search, you will be able to navigate to individual dataset pages. For individual dataset pages (instead of aggregate pages for e.g., a Family) there are three buttons available:

- Report issue - allows you to report any issues with the dataset in question
- Source DwC-A - download the dataset as a Darwin Core-Archive file. This will provide all data tables as separate files within a zipped folder
- To mapper - this will open another browser with the data shown in the Mapper

If you searched for aggregate datasets (e.g., all Crustacea records, all records from OBIS-Canada, etc.), the `source DwC-A` button will not be available to you. To download these data subsets, you must click `to mapper` and then download the data from the Mapper as a CSV.

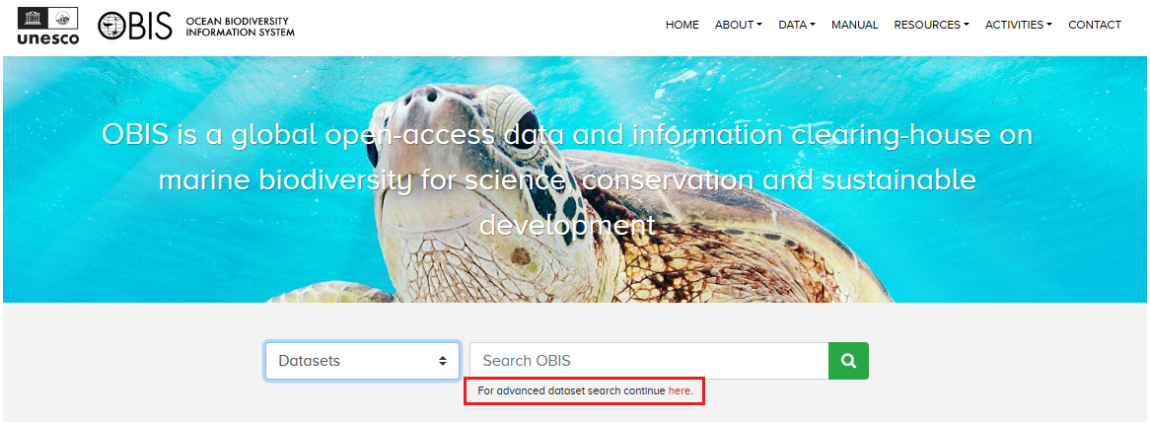


Figure 7.1: *OBIS homepage search, showing where to find the advanced search link*

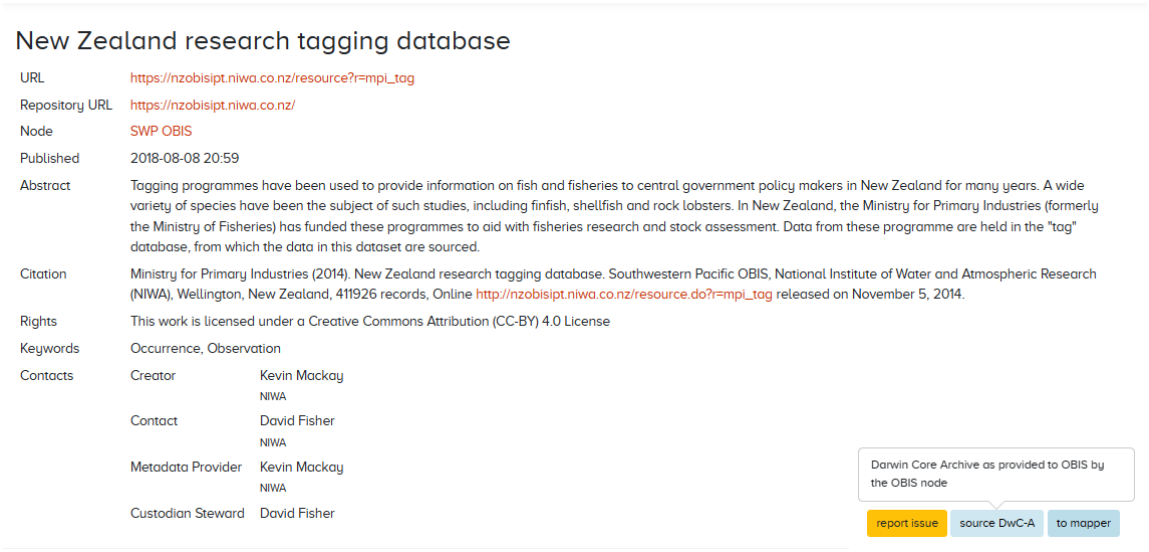


Figure 7.2: *Dataset download*

7.2 Mapper

- https://mapper.obis.org

Watch this video demonstration of how to use the Mapper as well as the OBIS homepage search.

The mapper allows users to visualize and inspect subsets of OBIS data. A variety of filters are available (taxonomic, geographic, time, data quality) and multiple layers can be combined in a single view. Layers can be downloaded as CSV files.

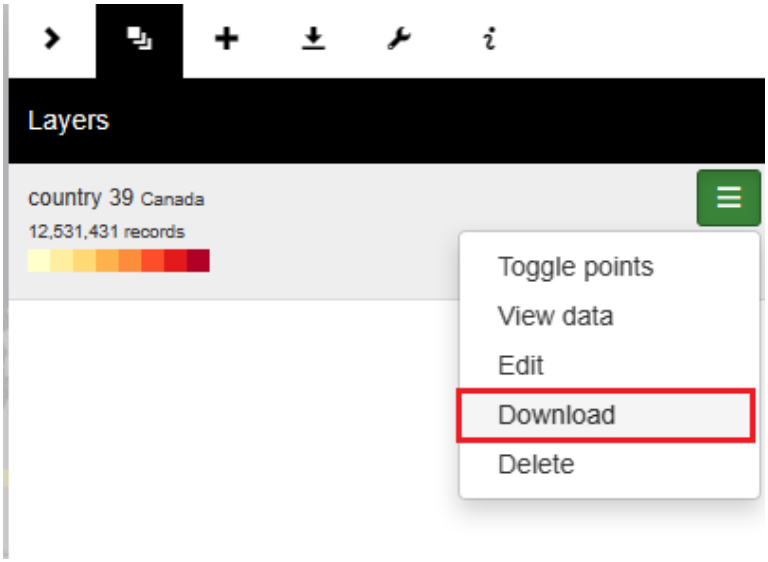


Figure 7.3: *Screenshot demonstrating where how to download a particular layer*

When you download data from the mapper, you will be given the option to include eMoF and/or DNA Derived Data extensions alongside the Event and Occurrence data. You must check the boxes of extensions you want to include in your download.

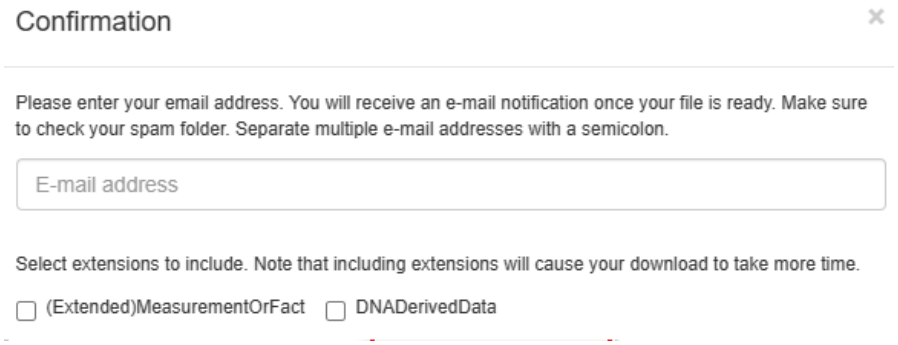


Figure 7.4: *Screenshot showing the popup confirmation for which extensions you want to include in your download from the OBIS Mapper*

After downloading, you will notice that the Event and Occurrence data is flattened into one table, called "Occurrence.csv". Upon inspecting this file in your viewer of choice, you will see it contains all 225 possible DwC fields, although not every field will contain data for each observation. Any extensions you checked will be downloaded as separate tables.

7.3 R package

- https://github.com/iobis/robis

The robis R package has been developed to facilitate connecting to the OBIS API from R. The package can be installed from CRAN or from GitHub (latest development version). The package documentation includes a function reference as well as a getting started vignette. As a quick example of what the package can do, you can obtain raw occurrence data by feeding a taxon name or AphiaID to the `occurrence` function.

If you'd like to then download this data, you can simply export R objects with the `write.csv` function. For example, if we wanted to obtain Mollusc data from OBIS:

```
library(robis)
moll<-occurrence("Mollusca")
write.csv(moll, "mollusca-obis.csv")
```

This file will be saved to your working directory (if you are not familiar with working directories, read here). After opening the file, you will notice that the fields in the download do not include every possible field, but instead only those where information has been recorded by data providers, plus the fields added by OBIS's quality control pipeline.

To use `robis` for visualizing and mapping occurrences, see the Visualization section of the manual.

Watch the video below for a walkthrough of how to use the robis package to obtain OBIS data.

7.4 API

- https://api.obis.org/

Both the mapper and the R package are based on the OBIS API, which can also be used to find and download data. When using the API directly, you can filter by the following options:

- Occurrence
- Taxon
- Checklist
- Node
- Dataset
- Institute
- Area
- Country
- Facet
- Statistics

When you have entered all the information you are interested in filtering by, scroll down and click the "Execute" button. This will produce a response detailing how many records match your criteria, as well as information for some of the headers from the data (e.g., basisOfRecord, Order, genus, etc.). A download button will be available for you to download the data as well.

When searching with the API, you may need to know certain identifiers, including:

- AphiaID - obtainable from the WoRMS page of a taxa of interest (e.g. the AphiaID for Mollusca would be 51)
- Dataset UUID - can be obtained from the URL on individual dataset pages
  - E.g., this dataset's UUID would be 5061d21c-6161-4ea2-a8d4-38f8285dfc47
- Area ID
- Institute ID - this should be the Ocean Expert ID (e.g., the ID for NOAA Fisheries Service, Southeast Regional Office St. Petersburg is 7532)
- OBIS node UUID

A short video demonstrating use of the API is shown below.

7.5 Full exports

- https://obis.org/data/access/

To obtain a full export of OBIS data, navigate to the OBIS homepage, click on Data from the top navigation bar, then select Data Access from the dropdown menu.

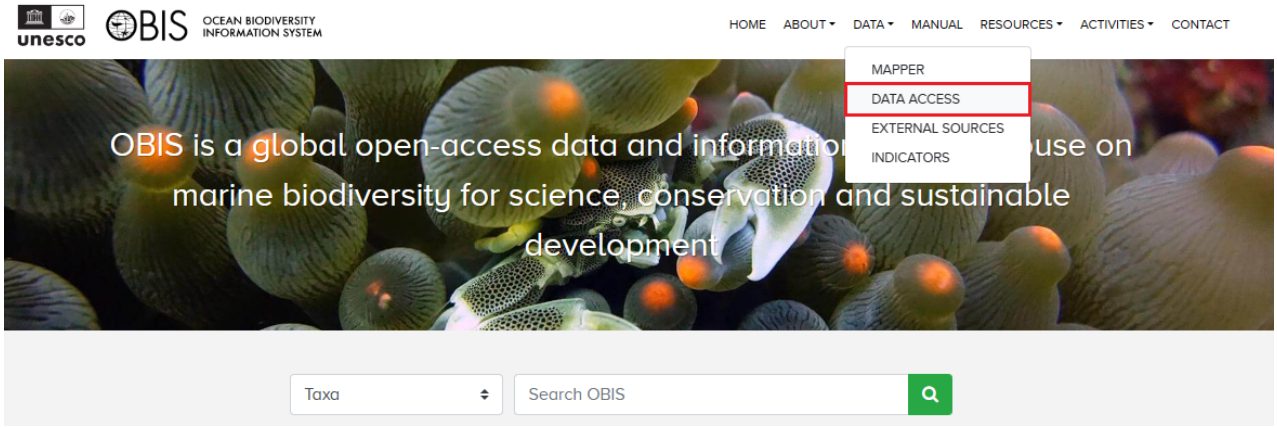


Figure 7.5: *OBIS homepage showing where to navigate to access full database exports*

Here you will be able to download all occurrence records as a CSV or Parquet file. Note the disclaimer that such exports will not include measurement data, dropped records, or absence records. As with downloads from the Mapper, the exported file is a single Occurrence table. This table includes all provided Event and Occurrence data, as well as 68 fields added by the OBIS Quality Control Pipeline, including taxonomic information obtained from WoRMS.

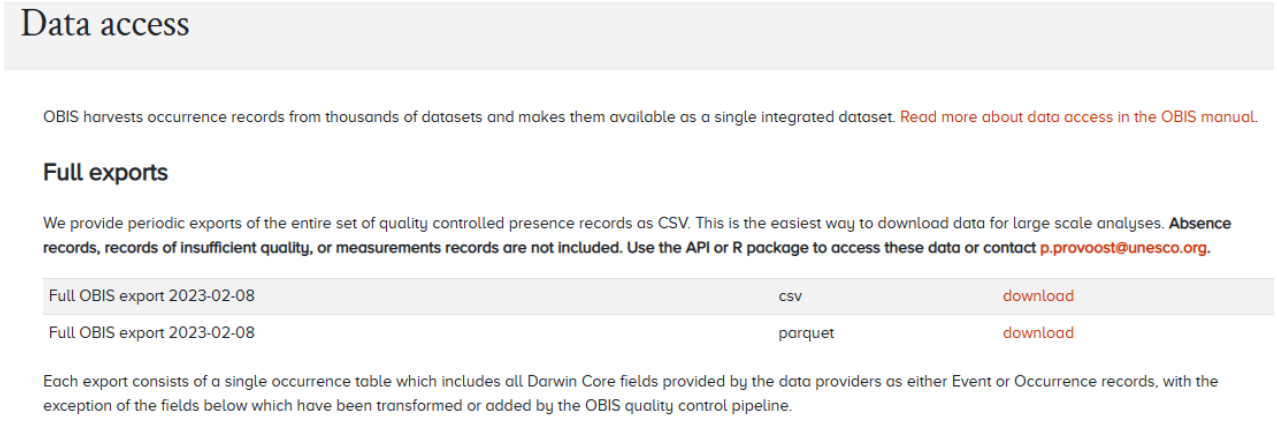


Figure 7.6: *OBIS Data Access page*

7.6 Finding your own data in OBIS

To find your own dataset in OBIS, you can use the same tools as finding any dataset in OBIS. You have the following options:

- From the OBIS homepage or the Mapper, you can search by dataset name, species of interest, the OBIS node that you uploaded to, or by institute
  - Note: When using the Mapper you can combine multiple search criteria to help narrow down your search
    * E.g., if we wanted to find this dataset in the Mapper, we could search for OBIS USA under Nodes, National Oceanic and Atmospheric Administration, Washington under Institutes, and/or Radiozoa under Scientific Name. Then when we view the data and scroll down to datasets, the only one listed is the one we were interested in
- If you have used the (extended)measurementOrFact extension and have measurementType data, you can search by the name of your measurementType, click on the hyperlink for records. This will populate a list of datasets that you can scroll through.

7.7 How to contact data provider

To contact the data provider, navigate to the page for the individual dataset in question (e.g., https://obis.org/dataset/80479e14-2730-436d-acaa-b63bdc7dd06f). Under the "Contacts" section, there will be a list of individuals you can contact. Clicking any name will direct you to your system's default email program. For example:

Figure 7.7: *Example of contact section on a dataset homepage access via the OBIS search*

If you are the node manager and need to contact the data provider about a particular dataset, contact information should be provided in the metadata and you can contact them from information provided.

7.8 Interpreting downloaded files from OBIS

In general, the field names you will see when you download data from OBIS are the same as those seen during the data formatting and publishing process. When you download data from the Mapper you will see all 225 possible Darwin Core fields.

Downloading data from an IPT or full export will include only the fields provided by the data provider, formatted as one Occurrence file (or separate files for individual datasets). Some fields are added through the OBIS quality control pipeline, including taxonomic information from WoRMS and the fields `flags`, `bathymetry`, and `dropped`. As mentioned in the Quality Control section, the fields `flags` and `dropped` will list quality control issues or if the record was dropped, respectively. Details and definitions for all fields added by the OBIS QC pipeline can be found here.

For a full list of the other Darwin Core terms and their definitions included in downloads, please reference the Darwin Core reference guide.

7.9 Citing Data from OBIS

Depending on how you access data from OBIS, there are different ways you should cite downloaded datasets.

General OBIS citation:

```
OBIS (YEAR) Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. www.obis.org.
```

For individual datasets retrieved from OBIS (dataset citations are available in the zip downloads as html file):

```
[Dataset citation available from metadata] [Data provider details] [Dataset] (Available: Ocean Biodiversity Information System. Intergovernmental O
```

For example:

```
Sousa Pinto, I., Viera, R. (Year: if not provided use year from dataset publication date) Monitoring of the intertidal biodiversity of rocky beache
```

When data represents a subset of many datasets taken from the integrated OBIS database (e.g. downloaded from the Mapper), you can, in addition to citing the individual datasets (and taking into account the restrictions set at each dataset level), also cite the OBIS database as follows:

```
OBIS (YEAR) [Data e.g. Distribution records of Eledone cirrhosa (Lamarck, 1798)] [Dataset] (Available: Ocean Biodiversity Information System. Inter
```

The derived information products from OBIS are published under the CC-0 license and can be cited as follows:

```
OBIS (YEAR) [Information product e.g. Global map showing the Hulbert index in a gridded view of hexagonal cells] [Map] (Available: Ocean Biodiversi
```

Data Visualization and Analysis

Chapter 8

Data Visualization

8.1 Example notebooks using data from OBIS

Here are a few R notebooks showcasing the robis package:

- Data exploration of wind farm monitoring datasets in OBIS
- Diversity of fish and vulnerable species in Marine World Heritage Sites based on OBIS data
- Data exploration - Stratified random surveys (StRS) of reef fish in the U.S. Pacific Islands
- DNADerivedData extension data access
- Canary Current LME

Here are others that may be of interest:

- Diversity indicators using OBIS data
- OBIS species richness for OSPAR
- Quality control of ISA data
- Accessing gridded data

8.2 obisindicators: calculating & visualizing spatial biodiversity using data from OBIS

obisindicators is an R library developed during the 2022 IOOS Code Sprint. The purpose was to create an ES50 diversity index within hexagonal grids following the diversity indicators notebook by Pieter Provoost linked above. The package includes several examples, limited to 1M occurrences, that demonstrate uses of the package.

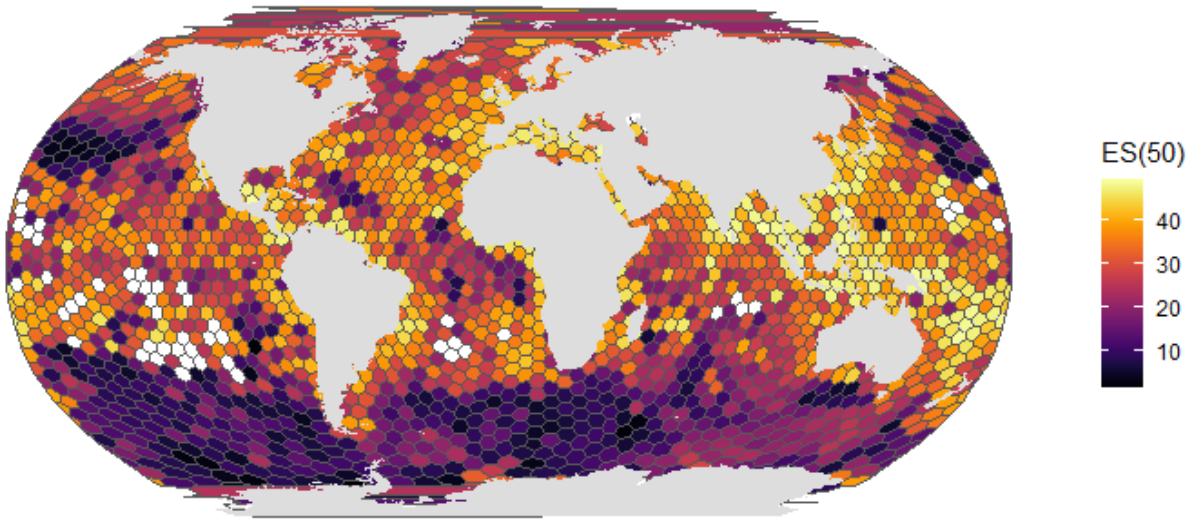


Figure 8.1: screenshot

Additional Resources

Chapter 9

Other Resources

In this section we highlight useful resources created by collaborators and other community members.

9.1 MBON Pole to Pole Tutorial

- https://www.youtube.com/watch?v=teJhfsSWonE

This tutorial was created by the MBON Pole to Pole project to help guide people through the process of transforming datasets to Darwin Core using tools MBON Pole to Pole has developed.

9.2 IOOS Darwin Core Guide

- https://ioos.github.io/bio_data_guide/

This book contains a collection of examples and resources related to mobilizing marine biological data to the Darwin Core standard for sharing though OBIS. This book has been developed by the Standardizing Marine Biological Data Working Group (SMBD). The working group is an open community of practitioners, experts, and scientists looking to learn and educate the community on standardizing and sharing marine biological data.

9.3 EMODnet Biology

- https://classroom.oceanteacher.org/course/view.php?id=430

Contributing Datasets to EMODnet Biology is a course hosted on Ocean Teacher Global Academy (OTGA), developed by members of the European Marine Observation and Data Network. The course prepares users to format, publish, and perform quality control checks on datasets according to Darwin Core standards. While targeted at EMODnet Biology users, this course has significant overlap in how to prepare datasets for OBIS and is useful for those unfamiliar with OBIS standards. Note, an account with OTGA is required to access the course.

9.4 Template Generators

There is an Excel template generator developed by Luke Marsden & Olaf Schneider as part of the Nansen Legacy project. It allows the creation of Event or Occurrence core templates, with an optional eMoF extension. Note this template generator is aimed at GBIF users, so make sure to account for and include required OBIS terms.

There is also an Excel to Darwin Core macro tool developed by GBIF Norway that you can download for use in Microsoft Excel. This macro can help you set up Event, Occurrence, and eMoF tables by selecting all relevant DwC fields from a list, or by importing data from another spreadsheet. It allows for auto-generation of identifiers (e.g. eventID, occurrenceID) if macros are enabled, and can also auto-populate the eMoF when measurement fields in the Occurrence table are populated.