The OBIS manual

20 March, 2023

Contents

Overview **6**

1 Introduction **6**
1.1 Guidelines on the sharing and use of data in OBIS . . . 6
1.2 Acknowledgements . . . 6
1.3 Data Policy . . . 6
1.4 Getting Help in OBIS . . . 8

Contributing data to OBIS **10**

2 What can you contribute and how? **10**
2.1 Why publish data to OBIS . . . 11
2.2 How to handle sensitive data . . . 12
2.3 OBIS Data Life Cycle . . . 12
2.4 Biodiversity data standards . . . 13
2.5 OBIS nodes . . . 30

Data Formatting **34**

3 Dataset structure **34**
3.1 When to use Event Core . . . 34
3.2 When to use Occurrence Core . . . 34
3.3 Extensions in OBIS . . . 35
3.4 Constructing and using indentifier codes . . . 35

4 Formatting data tables **38**
4.1 Darwin Core Term Checklist for OBIS . . . 38
4.2 Name Matching Strategy for taxonomic quality control . . . 41
4.3 How to format Occurrence tables . . . 44
4.4 How to format Event tables . . . 45
4.5 How to format extendedMeasurementOrFact tables . . . 46
4.6 Choosing Vocabularies for your dataset . . . 46
4.7 Common Data formatting issues . . . 47
4.8 Examples: ENV-DATA and DNA derived data . . . 49

Ensuring Data Quality **67**

5 Data quality control **67**
5.1 Data quality flags . . . 68

5.2 Common Quality Control issues 71

Publishing Data **73**

6 Data publication and sharing **73**
6.1 Add metadata 73
6.2 Licenses 73
6.3 IPT Administration 73
6.4 IPT 73
6.5 Maintaining published data 77
6.6 Simultaneous publishing to GBIF 77
6.7 Update your data in OBIS 77

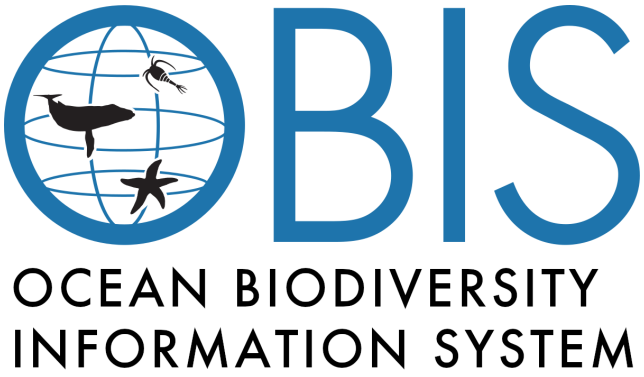
Access Data from OBIS **79**

7 Data access **79**
7.1 Mapper 79
7.2 R package 79
7.3 API 79
7.4 Full exports 79
7.5 Finding your own data 79
7.6 Citing Data from OBIS 79

Data Visualization and Analysis **82**
7.7 Example notebooks using data from OBIS 82
7.8 obisindicators: calculating & visualizing spatial biodiversity using data from OBIS 82

Additional Resources **84**

8 Other Resources **84**
8.1 MBON Pole to Pole Tutorial 84
8.2 IOOS Darwin Core Guide 84
8.3 EMODnet Biology 84


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 and Ward Appeltans.

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
- CC-BY
- CC-BY-NC

CC-0 is the preferred one and CC-BY-NC the 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

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 so you are more likely to receive a quick answer to your question by posting there, as the Secretariat receives many requests.

Finally, you can submit an issue on relevant Github repositories:

- OBIS Manual
- OBIS Website
- OBIS issues GitHub repo
- All other OBIS repositories

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. [LINK TO TIPS FOR QC?]
5. Now that your dataset is ready for publishing, the relevant metadata must be filled in, and then published on the previously identified IPT[link to IPT topic section].

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[LINK to access section], 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. Maptool tutorial 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

- Introduction to Darwin Core
- Darwin Core terms

- Darwin Core guidelines
  - Taxonomy and identification
  - Occurrence

  - Record level terms

  - Location

  - Event
  - Time

  - Sampling

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)](http://tdwg.org/). 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 eight required DwC terms: `occurrenceID`, `eventDate`, `decimalLongitude`, `decimalLatitude`, `scientificName`, `scientificNameID`, `occurrenceStatus`, `basisOfRecord`.

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
- 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 synomym. 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` (required term), OBIS will use this identifier to pull the taxonomic information from the World Register of Marine Species (WoRMS) into OBIS, such as the taxonomic classification and the accepted name in case of invalid names or synonyms. 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 uncertain 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 uncertain species name combined with *?* or other qualifiers. The table belowe shows a few examples:

The use and definitions for additional NO 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
Pelagia	Péron & Lesueur, 1810	urn:lsid:marinespecies.org:taxname:135262	genus	gen. nov.
Pelagia benovici	Piraino, Aglieri, Scorrano & Boero, 2014	urn:lsid:marinespecies.org:taxname:851656	species	sp. nov
Gadus	Linnaeus, 1758	urn:lsid:marinespecies.org:taxname:125732	genus	cf. morhua
Polycera	Cuvier, 1816	urn:lsid:marinespecies.org:taxname:138369	genus	cf. hedgpethi
Tubifex	Lamarck, 1816	urn:lsid:marinespecies.org:taxname:137392	genus	?
Tubifex	Lamarck, 1816	urn:lsid:marinespecies.org:taxname:137392	genus	sp. inc.
Brisinga	Asbjørnsen, 1856	urn:lsid:marinespecies.org:taxname:123210	genus	gen. inc.
Uroptychus compressus	Baba & Wicksten, 2019	urn:lsid:marinespecies.org:taxname:1332465	genus	sp. inc.
Eurythenes	S. I. Smith in Scudder, 1882	urn:lsid:marinespecies.org:taxname:101607	genus	sp. DISCOLL.PAP.JC165.674
Paroriza	Hérouard, 1902	urn:lsid:marinespecies.org:taxname:123467	genus	sp.[unique123]aff.pallens
Aristeidae	Wood-Mason in Wood-Mason & Alcock, 1891	urn:lsid:marinespecies.org:taxname:106725	family	stet.
Nematocarcinus	Milne-Edwards, 1881	urn:lsid:marinespecies.org:taxname:107015	genus	sp.indet.
Brisinga	Asbjørnsen, 1856	urn:lsid:marinespecies.org:taxname:123210	genus	gen.inc.
Brisinga costata	Verrill, 1884	urn:lsid:marinespecies.org:taxname:17825	species	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. There are no guidelines yet on designing the persistence of this ID, the level of uniqueness (from dataset to global) and the precise algorithm and format for generating the ID, but in the absence of a persistent globally unique identifier, one could be constructed by combining the `institutionCode`, the `collectionCode` and the `catalogNumber` (or autonumber in the absence of a catalogNumber), see further below. Note that the inclusion of occurrenceID is also necessary for datasets in the OBIS-ENV-DATA format.

`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 treated 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` (fossil, which allows OBIS to make the distinction between the date of collection and the time period the specimen was assumed alive) or `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 DNA sequencers, image recognition etc).

When the basisOfRecord is 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 `type` (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 `catalogNumber` is an identifier for the record within the dataset or collection.

`occurrenceID` is detailed in Identifiers

`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)))
```

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.

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.

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. OBIS recommends using the extended ISO 8601 format with hyphens.

ISO 8601 dates can represent moments in time at different resolutions, as well as time intervals, which use `/` as a separator. Date and time are separated by `T`. Times can have a time zone indicator at the end, if this is not the case then the time is assumed to be local time. When a time is UTC, a `Z` is added. Some examples of ISO 8601 dates are:

```
1973-02-28T15:25:00
2005-08-31T12:11+12
1993-01-26T04:39+12/1993-01-26T05:48+12
2008-04-25T09:53
1948-09-13
1993-01/02
1993-01
1993
```

Besides year, month and day numbers, ISO 8601 also supports ordinal dates (year and day number within that year) and week dates (year, week, and day number within that week). These dates are less common and have the formats `YYYY-DDD` (for example `2015-023`) and `YYYY-Www-D` (for example `2014-W26-3`).

ISO 8601 durations should not be used.

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

- Darwin Core Archive
- OBIS holds more than just species occurrences: the ENV-DATA approach
  - ExtendedMeasurementOrFact Extension (eMoF)
  - eDNA & DNA derived data Extension
  - A special case: habitat types
- When to use Event Core
- When to use Occurrence Core
- Recommended reading

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.

2.4.2.3 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 part of the DwC standardized vocabularies and are integral to reducing repeated

data, especially when one dataset has multiple years. 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 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.

Figure. 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.

2.4.2.4 How to avoid redundancy

Avoiding redundancy and data duplication within your dataset is built into the OBIS data structure. Utilizing the star schema 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.

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. A possible solution to avoid duplicating Event data in these cases is to publish Event data as sibling datasets. Thus you would have two datasets, linked by the eventID:

- 1 Event core + eMoF
- 1 Occurrence core + DNA-derived data extensions + eMoF

2.4.3 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: `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.

2.4.3.1 Scenarios

2.4.3.1.1 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 titel 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 increae 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 elsewere, 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.

:exclamation: 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

2.4.3.1.2 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.

2.4.3.1.3 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', 'principle 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 principle contact/creator should, however, be added in the IPT Basic Metadata 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)
- Principle 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".

2.4.3.1.4 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 at 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 commericial use is not allowed. This licence can be problematic when the data is re-used in scientific journals.

2.4.3.1.5 Coverage

2.4.3.1.5.1 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.

2.4.3.1.5.2 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.

2.4.3.1.5.3 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.

2.4.3.1.6 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.

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.

2.4.3.1.7 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). This is not a problem.

2.4.3.1.8 Sampling Methods The EML can contains descriptions of the sampling and data processing methods. Note that OBIS best practice is to add sampling facts to the extended MeasurementorFact extension, linked to the sampling events in the EventCore.

2.4.3.1.9 Citations The dataset citation allows users to properly cite the datasets in further publications or other uses of the data. The OBIS download function provides a list of the dataset citations packaged with the data in a zipped file. 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 citaiton, 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}

2.4.3.1.10 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.

2.4.3.1.11 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.

2.4.3.1.12 External Links This section can include URLs to the resource homepage, to download or find additional information.

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.

2.4.3.1.13 Additional Metadata Any remaining information that could not be catalogued under any of the other metadata, can be mentioned here.

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.

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 no specific actions 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.

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. 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 newly developed 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 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 of 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

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.

Differences between OBIS Core Table Identifiers

If your core file is based on occurrences (e.g., an observation), then any extensions are linked with `occurrenceID`. If your core file is based on events (e.g., a sampling event), then the linking identifier is `eventID`. Thus, 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.

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 Constructing and using indentifier codes

3.4.1 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 eventIDs 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 parentID 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 eventIDs to ensure codes remain globally unique in subsequent years, which is particularly useful if your sampling protocol is repeated temporally.

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	type	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	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 eventIDs 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 eventIDs 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.

3.4.2 occurrenceID

`occurrenceID` is an identifier for occurrence records. Each occurrence record must have a 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 exist.

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. 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	required	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

Term	OBIS Required	DarwinCore Class	Event	Occurrence	eMoF	DNA
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
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		

Term	OBIS Required	DarwinCore Class	Event	Occurrence	eMoF	DNA
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		
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		

Term	OBIS Required	DarwinCore Class	Event	Occurrence	eMoF	DNA
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 species names to be matched against an authoritative taxonomic register. This effectively attaches stable identifiers 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.

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. 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 which indicate the authority the ID comes from.

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:

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 list of your species
- Upload to WoRMS taxon match tool
  - Check relevant boxes
- Review returned file
- Identify data to include for OBIS
  - LSIDs, taxonomic fields, etc.
- Attach LSIDs back to your data using either:
  - R (merge)
  - Excel (vlookup)

The taxon match tool of the World Register of Marine Species (WoRMS) 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.

This 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.

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.

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. We have also created a [video tutorial] for using the taxon match tool, including how to attach the IDs 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 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

4.2.1.3 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 unchcek 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

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

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

4.2.3.1 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.

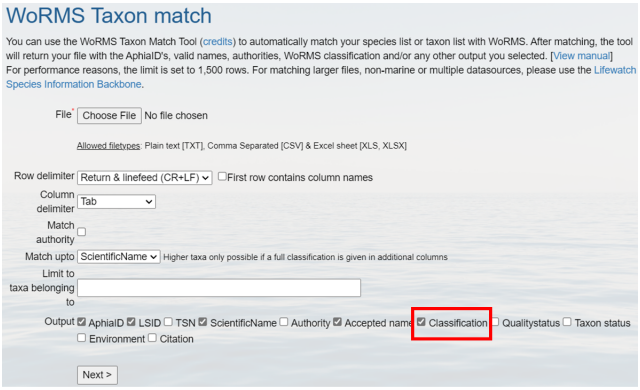


Figure 4.1: WoRMS classification box

4.2.3.2 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 :

- 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.

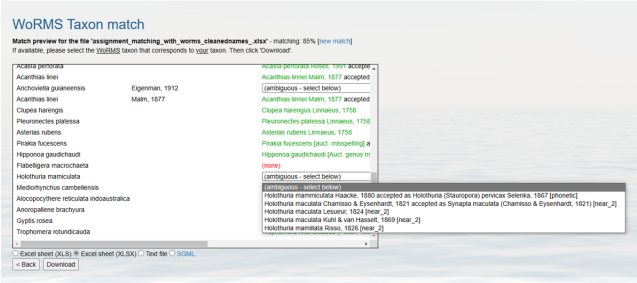


Figure 4.2: Example of choices from an abiguous match

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

- First 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
- 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, you should contact info@marinespecies.org to see if the taxon should be added to the WoRMS register.

4.3 How to format Occurrence tables

If your dataset structure is based on Occurrence core, or has an Occurrence extension, 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`
- `eventDate` (not required for Occurrence extension, required for Occurrence Core)
- `decimalLatitude` (not required for Occurrence extension)
- `decimalLongitude` (not required for Occurrence extension)

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: scientificNameID 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.

4.3.1 Stepwise Guidance to Format an Occurrence Table (in Excel)

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 added to the eMoF table
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.

4.4.1 Stepwise Guidance to Format Event Table (in Excel)

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

1. Identify 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.

2. Copy these columns to a new sheet and name it Event
3. Delete duplicate data so only unique events are left.
4. Identify the hierarchical event structure in your data, if present
5. Add and fill the `parentEventID` and `type` fields as applicable
6. Create new records for parent Events
7. Ensure dates and time are formatted according to ISO 8601 standards in the eventDate field
8. Add any other relevant fields as indicated above

Watch the video tutorial of this process.

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.

4.5 How to format extendedMeasurementOrFact tables

4.6 Choosing Vocabularies for your dataset

4.6.1 Match your data with DwC vocabulary

4.6.2 How to correctly map eMoF terms to preferred BODC vocabulary

4.6.2.0.1 MeasurementOrFact vocabularies The MeasurementOrFact terms `measurementType`, `measurementValue` and `measurementUnit` are completely unconstrained and can be populated with free text annotation. While free text offers the advantage of capturing complex and as yet unclassified information, the inevitable semantic heterogeneity (e.g. of spelling or wording) becomes a major challenge for effective data integration and analysis. Hence, OBIS added 3 new terms: `measurementTypeID`, `measurementValueID` and `measurementUnitID` to standardise the measurement types, values and units. Note that `measurementValueID` is not used for standardizing numeric measurements. The three new terms should be populated using controlled vocabularies referenced using Unique Resource Identifiers (URIs). OBIS recommends to use the internationally recognized NERC Vocabulary Server, developed by the British Oceanographic Data Centre (BODC), which can be searched through https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/.

The following vocabularies are recommended for populating `measurementTypeID`, `measurementValueID`, and `measurementUnitID`:

4.6.2.0.1.1 measurementTypeID

- BODC Parameter Usage Vocabulary (P01)
  - documentation: https://github.com/nvs-vocabs/P01
  - vocabulary: http://vocab.nerc.ac.uk/collection/P01/current/
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/P01/
- OBIS sampling instruments and methods attributes (Q01)
  - vocabulary: http://vocab.nerc.ac.uk/collection/Q01/current/
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/Q01/

4.6.2.0.1.2 measurementValueID

- Sampling instruments and sensors (SeaVoX Device Catalogue)
  - documentation: https://github.com/nvs-vocabs/L22
  - vocabulary: http://vocab.nerc.ac.uk/collection/L22/current
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/L22/
- Sampling instrument categories (SeaDataNet device categories)
  - documentation: https://github.com/nvs-vocabs/L05

  - vocabulary: http://vocab.nerc.ac.uk/collection/L05/current
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/L05/
- Vessels (ICES Platform Codes)
  - vocabulary: http://vocab.nerc.ac.uk/collection/C17/current
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/C17/
- Sex (Gender)
  - documentation: https://github.com/nvs-vocabs/S10
  - vocabulary: http://vocab.nerc.ac.uk/collection/S10/current/
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/S10/

- Lifestage
  - documentation: https://github.com/nvs-vocabs/S11
  - vocabulary: http://vocab.nerc.ac.uk/collection/S11/current/
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/S11/
- Papers or manuals on the sampling protocol used
  - DOI
  - Handle for publications on IOC's Ocean Best Practices repository, for example: http://hdl.handle.net/11329/304

4.6.2.0.1.3 MeasurementUnitID

- Units
  - documentation: https://github.com/nvs-vocabs/P06
  - vocabulary: http://vocab.nerc.ac.uk/collection/P06/current
  - search: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/P06/

4.6.3 Requesting new vocabulary terms

4.7 Common Data formatting issues

4.7.1 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 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.

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

- **`scientificNameID`**

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 WoRMS and other LSIDs, 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 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, DNA sequences)

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.

4.7.2 Temporal: Dates and times

4.7.2.1 Historical data

4.7.3 Spatial

4.7.3.1 Converting Coordinates

4.7.3.2 Geographical formats

4.8 Examples: ENV-DATA and DNA derived data

Contents

- Fish abundance & distribution
- Hard coral cover & composition
- Invertebrates abundance & distribution
- Macroalgae canopy cover & composition
- Mangroves cover & composition
- 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

4.8.0.1 Fish abundance & distribution

(example coming soon)

4.8.0.2 Hard coral cover & composition

(example coming soon)

4.8.0.3 Invertebrates abundance & distribution

(example coming soon)

4.8.0.4 Macroalgae canopy cover & 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/AREALDS	0.25		m2	P06/current/UMSQ
quadrat_1	occ_1	cover	P01/current/SDBIOL	24	0	percent	P06/current/UPCT
quadrat_1	occ_2	cover	P01/current/SDBIOL	56	0	percent	P06/current/UPCT
transect_1		surface area	P01/current/AREALDS	60		m2	P06/current/UMSQ
transect_1		length	P01/current/LENTRCK	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/SDBIOL	8	10	percent	P06/current/UPCT
transect_1	occ_4	cover	P01/current/SDBIOL	24	0	percent	P06/current/UPCT
transect_2	occ_5	cover	P01/current/SDBIOL	4	10	percent	P06/current/UPCT
transect_2	occ_6	cover	P01/current/SDBIOL	16	0	percent	P06/current/UPCT

4.8.0.5 Mangroves cover & composition

(example coming soon)

4.8.0.6 Marine birds abundance & 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
in2017_v02_00998	2017-03-17 01:07:00	Australasian Seabird Group, BirdLife Australia	in2017_v02_wov
in2017_v02_01380	2017-03-19 22:26:00	Australasian Seabird Group, BirdLife Australia	in2017_v02_wov
in2017_v02_01012	2017-03-17 02:38:00	Australasian Seabird Group, BirdLife Australia	in2017_v02_wov

basisOfRecord	recordedBy	organismQuantity	organismQuantityType	occurrenceStatus
HumanObservation	EJW+CRC+TAH	2	individuals	present
HumanObservation	EJW+CRC+TAH	1	individuals	present
HumanObservation	EJW+CRC+TAH	1	individuals	present

decimalLatitude	decimalLongitude	coordinateUncertaintyInMeters	coordinatePrecision	footprintWKT
-43.40741	147.45576	200	0.0018	POINT (147.45576 -43.40741)
-45.98644	142.1445	200	0.0018	POINT (142.14450 -45.98644)
-43.40728	147.45549	200	0.0018	POINT (147.45549 -43.40728)

scientificNameID	scientificName	scientificNameAuthorship	vernacularName
urn:lsid:marinespecies.org:taxname:343991	Morus serrator	(Gray,1843)	Australasian Gannet
urn:lsid:marinespecies.org:taxname:212648	Pachyptila turtur	(Kuhl,1820)	Fairy Prion
urn:lsid:marinespecies.org:taxname:707545	Chroicocephalus novaehollandiae	Stephens,1826	Silver Gull

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
in2017_v02_00998	in2017_v02_00998-depth	Sea-floor depth (m)	http://vocab.aodn.org.au/def/discovery_parameter/entity/574
in2017_v02_00998	in2017_v02_00998-depth	Sea-floor depth	http://vocab.nerc.ac.uk/collection/P01/current/MBANZZZZ/
in2017_v02_00998	in2017_v02_00998-air_pressure	Air Pressure (hPa)	http://vocab.nerc.ac.uk/collection/P01/current/CAPHZZ01
in2017_v02_00998	in2017_v02_00998-air_temp	Atmospheric temperature (deg C)	http://vocab.nerc.ac.uk/collection/P01/current/CTMPZZ01
in2017_v02_00998	in2017_v02_00998-wov_sea_state	Sea state	http://vocab.nerc.ac.uk/collection/C39/current/
in2017_v02_00998	in2017_v02_00998-sea_surface_temp	Sea surface temperature	http://vocab.aodn.org.au/def/discovery_parameter/entity/97
in2017_v02_00998	in2017_v02_00998-sea_surface_temp	Sea surface temperature	http://vocab.nerc.ac.uk/standard_name/sea_surface_temperature/
in2017_v02_00998	in2017_v02_00998-wind_direction	Wind direction (deg)	http://vocab.nerc.ac.uk/collection/P01/current/EWDAZZ01

occurrenceID	measurementID	measurementType	measurementTypeID
in2017_v02_00998	in2017_v02_00998-wind_speed	Wind Speed (knt)	http://vocab.nerc.ac.uk/collection/P01/current/ESSAZZ01

measurementValue	measurementValueID	measurementUnit	measurementUnitID
73.0313	NA	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA
73.0313	NA	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA
1024.91385	NA	hPa	http://vocab.nerc.ac.uk/collection/P06/current/HPAX
15.3	NA	degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA
moderate 1.25 - 2.5 m	http://vocab.nerc.ac.uk/collection/C39/current/4/		
17.32	NA	degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA
17.32	NA	degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA
283	NA	degrees	http://vocab.nerc.ac.uk/collection/P06/current/UABB
5.49	NA	Knots (nautical miles per hour)	http://vocab.nerc.ac.uk/collection/P06/current/UKNT

4.8.0.7 Marine mammals abundance & 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

4.8.0.8 Marine turtles abundance & distribution

4.8.0.8.1 Survey & 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
TURTLE_CBAR_2013	2014-05-05 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_2015	2014-05-05 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_0045_0001	2014-05-05 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_0045_0002	2014-05-05 07:59:08	https://marineinfo.org/id/dataset/6403	Presence of sea turtles collected through Fixed-Line-Transect monitoring across the Western Mediterranean Sea

eventID	parentEventID	eventDate
TURTLE_CBAR_0043		2013-04-03T05:30:00+02:00/2013-04-03T16:00:00+02:00
TURTLE_CBAR_0045		2013-04-18T05:22:00+02:00/2013-04-18T15:53:00+02:00
TURTLE_CBAR_0045_0001	TURTLE_CBAR_0045	2013-04-18T05:55:00+02:00
TURTLE_CBAR_0045_0002	TURTLE_CBAR_0045	2013-04-18T08:35:00+02:00

eventRemarks	minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude
transect	0	0	41.26179967	4.933265167
transect	0	0	41.30371367	4.936571167
sample	0	0	41.3228	7.4984

eventRemarks	minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude
sample	0	0	41.322845	5.995345

geodeticDatum	coordinateUncertaintyInMeters	footprintWKT	footprintSRS
EPSG:4326	222970.2874	LINESTRING (7.602633333333 41.24378333333, 2.263897 41.279816)	EPSG:4326
EPSG:4326	225420.0359	LINESTRING (7.636983333333 41.32418333333, 2.236159 41.283244)	EPSG:4326
EPSG:4326		POINT	EPSG:4326
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
TURTLE_CBAR_0043	AdL_TURTLE_CBAR_0001	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_13-17	HumanObservation
TURTLE_CBAR_0045	AdL_TURTLE_CBAR_0004	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_13-17	HumanObservation
TURTLE_CBAR_0045_0001	AdL_TURTLE_CBAR_0005	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_13-17	HumanObservation
TURTLE_CBAR_0045_0002	AdL_TURTLE_CBAR_0006	https://marineinfo.org/id/dataset/6403	TURTLE_CBAR_13-17	HumanObservation

catalogNumber	recordedBy	occurrenceStatus
AdL_TURTLE_CBAR_0001	Ilaria Campana | Miriam Paraboschi | Erica Ercoli | Erica	absent
AdL_TURTLE_CBAR_0004	Antonella Arcangeli | Cristina Berardi | Lucilla Giulietti | Claudia Boccardi	absent
AdL_TURTLE_CBAR_0005	Antonella Arcangeli | Cristina Berardi | Lucilla Giulietti | Claudia Boccardi	present
AdL_TURTLE_CBAR_0006	Antonella Arcangeli | Cristina Berardi | Lucilla Giulietti | Claudia Boccardi	present

scientificNameID	scientificName	kingdom	scientificNameAuthorship
urn:lsid:marinespecies.org:taxname:136999	Cheloniidae	Animalia	Oppel, 1811
urn:lsid:marinespecies.org:taxname:136999	Cheloniidae	Animalia	Oppel, 1811
urn:lsid:marinespecies.org:taxname:137205	Caretta caretta	Animalia	Linnaeus, 1758
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
TURTLE_CBAR_0043	AdL_TURTLE_CBAR_0001	BEAUFORT WIND FORCE	http://vocab.nerc.ac.uk/collection/P01/current/WMOCWFBF
TURTLE_CBAR_0045	AdL_TURTLE_CBAR_0004	Platform height	http://vocab.nerc.ac.uk/collection/P01/current/AHSLZZ01
TURTLE_CBAR_0045	AdL_TURTLE_CBAR_0004	Sampling method	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100003
TURTLE_CBAR_0045	AdL_TURTLE_CBAR_0004	Speed of measurement platform relative to ground surface {speed over ground} by unspecified GPS system	http://vocab.nerc.ac.uk/collection/P01/current/APSAGP01
TURTLE_CBAR_0045_0001	AdL_TURTLE_CBAR_0005	Development stage of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01
TURTLE_CBAR_0045_0001	AdL_TURTLE_CBAR_0005	Count of assayed sample) of biological entity specified elsewhere	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01
TURTLE_CBAR_0045_0001	AdL_TURTLE_CBAR_0005	Sighting distance	

measurementValue	measurementUnit	measurementUnitID
0	Beaufort scale	
29	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA/

measurementValue	measurementUnit	measurementUnitID
visual observation from ferries		
23.291	Knots (nautical miles per hour)	http://vocab.nerc.ac.uk/collection/P06/current/UKNT/
1		
20	Metres	http://vocab.nerc.ac.uk/collection/P06/current/ULAA/

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
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

4.8.0.9 Microbes biomass & diversity

(example coming soon)

4.8.0.10 Phytoplankton biomass & 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.

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	eventRemarks	eventDate	modified
TripNR3242		cruise	2017-05T13:18:00+00:00/2017-05T22:14:00+00:00	2021-10-21 15:52:00
TripNR3242TripStationNR16781	TripNR3242	stationVisit	2017-05-08T20:44:00+00:00/2017-05-08T20:55:00+00:00	2021-10-21 15:52:00
TripNR3242TripStationNR16781MidasTripActionID10558	TripNR3242TripStationNR16781	sample	2017-05-08T20:50:00+00:00	2021-10-21 15:52:00
TripNR3242TripStationNR16781MidasTripActionID10558	TripNR3242TripStationNR16781	sample	2017-05-08T20:50:00+00:00	2021-10-21 15:52:00

datasetID	datasetName	locationID	waterBody	country	countryCode
https://marineinfo.org/id/dataset/4688	LifeWatch observatory data: phytoplankton observations...		North Sea	Belgium	BE
https://marineinfo.org/id/dataset/4688		JN17_5			
https://marineinfo.org/id/dataset/4688		JN17_5			
https://marineinfo.org/id/dataset/4688		JN17_5			

minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	geodeticDatum	coordinateUncertaintyInMeters	footprintSRS
				EPSG:4326		EPSG:4326
0	30.22	51.0131	1.90562	EPSG:4326	1.11	EPSG:4326
0	1	51.01203	1.90217	EPSG:4326	1.11	EPSG:4326
3	3	51.01203	1.90217	EPSG:4326	1.11	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
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__(Pseudo-)pediastrum__5
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__Actinoptychus senarius__5
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__Actinoptychus splendens__5
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__Actinoptychus__5

modified	basisOfRecord	occurrenceStatus	scientificNameID	scientificName
2021-10-21	Occurrence	absent	urn:lsid:marinespecies.org:taxname:160560	Hydrodictyaceae
2021-10-21	Occurrence	present	urn:lsid:marinespecies.org:taxname:148948	Actinoptychus senarius
2021-10-21	Occurrence	present	urn:lsid:marinespecies.org:taxname:148949	Actinoptychus splendens
2021-10-21	Occurrence	present	urn:lsid:marinespecies.org:taxname:148947	Actinoptychus

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
TripNR3242		Platform Name
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__Actinoptychus	Abundance of Actinoptychus (WoRMS:148947) per unit volume of the water body by image analysis
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__Actinoptychus senarius__5	Abundance of Actinoptychus senarius (WoRMS:148948) per unit volume of the water body by image analysis
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__(Pseudo-)pediastrum__5	Lifestage
TripNR3242TripStationNR16781MidasTripActionID105598	TripNR3242TripStationNR16781MidasTripActionID105598occurenceIDTA__105598__Actinoptychus senarius__5	Lifestage
TripNR3242TripStationNR16781MidasTripActionID105598		Sampling device aperture diameter
TripNR3242TripStationNR16781MidasTripActionID105598		Sampling instrument name
TripNR3242TripStationNR16781MidasTripActionID105598		Sampling net mesh size
TripNR3242TripStationNR16781MidasTripActionID105598UW		Conductivity of the water body
TripNR3242TripStationNR16781MidasTripActionID105598UW		Practical salinity of the water body
TripNR3242TripStationNR16781MidasTripActionID105598UW		Temperature of the water body

measurementTypeID	measurementValue	measurementValueID	measurementUnit
http://vocab.nerc.ac.uk/collection/Q01/current/Q0100001/	Simon Stevin	http://vocab.nerc.ac.uk/collection/C17/current/11SS/	
http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/	2.24		specimens/L
http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/	1.12		specimens/L
http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01/	adult	http://vocab.nerc.ac.uk/collection/S11/current/S1116/	
http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01/	adult	http://vocab.nerc.ac.uk/collection/S11/current/S1116/	
http://vocab.nerc.ac.uk/collection/Q01/current/Q0100012/	0.4		meter

measurementTypeID	measurementValue	measurementValueID	measurementUnit
http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Planktonnet Apstein	http://vocab.nerc.ac.uk/collection/L22/current/TOOL0978/	
http://vocab.nerc.ac.uk/collection/Q01/current/Q0100015/	55		micrometer
http://vocab.nerc.ac.uk/collection/P01/current/CNDCZZ01/	3.916		Siemens per metre
http://vocab.nerc.ac.uk/collection/P01/current/PSALPR01/	34.295		Grams per kilogram
http://vocab.nerc.ac.uk/collection/P01/current/TEMPPR01/	11.881		Degrees Celsius

measurementUnitID	measurementDeterminedBy	measurementMethod
	Flanders Marine Institute	
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
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
	Flanders Marine Institute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human
	Flanders Marine Institute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human
http://vocab.nerc.ac.uk/collection/P06/current/ULAA/	Flanders Marine Institute	
	Flanders Marine Institute	
http://vocab.nerc.ac.uk/collection/P06/current/UMIC/	Flanders Marine Institute	
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
http://vocab.nerc.ac.uk/collection/P06/current/UGKG/	Flanders Marine Institute	Practical salinity of the water body based on water from the UnderWaySystem of the ship
http://vocab.nerc.ac.uk/collection/P06/current/UPAA/	Flanders Marine Institute	Temperature of the water body based on water from the UnderWaySystem of the ship

4.8.0.11 Seagrass cover & 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
USBsg-chengue-pastocoral		2019-05-13	seagrass	no notes	CO
USBsg-chengue-pastomanglar		2019-05-14	seagrass	no notes	CO
USBsg-chengue-pastocoral-SquidPopTransect1	USBsg-chengue-pastocoral	2019-05-13	seagrass	no notes	CO
USBsg-chengue-pastocoral-SquidPopTransect2	USBsg-chengue-pastocoral	2019-05-13	seagrass	no notes	CO

minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	coordinateUncertaintyInMeters	footprintWKT	footprintSRS
0.8	2	11.32021806	-74.12753684	10	POLYGON ((-74.1273259763024 11.320475512862,-74.1272978004008 11.3201655779439))	EPSG:4326
0.8	0.8	11.31977189	-74.12536879	10	POLYGON ((-74.1253370891273 11.3195001294432,-74.1253337743154 11.3194968146313))	EPSG:4326

minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	coordinateUncertaintyInMeters	footprintWKT	footprintSRS
0.8	2	11.32039927	-74.12737404	50	POINT (-74.1273740410759 11.3203992721869)	EPSG:4326
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
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-PhyQ01	WaterTemp
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-PhyQ02	Salinity
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-PhyQ03	Dissolved oxygen
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1C1-shoot-01	Shoot Density
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1C1-leafLenght-01	Leaf Length
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-DryBiomass	Total Dry Biomass
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-biomassGL	Dry biomass of green leaves
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-biomassNGL	Dry biomass of non green leaves
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N1-biomassSH	Dry biomass of the shoots
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N2-biomassR	Dry biomass of the roots
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N2-biomassRIZ	Dry biomass of the rizome
USBsg-chengue-pastocoral	USBsg-chengue-pastocoral-tt	USBsg-chengue-pastocoral-T1N2-biomassOTH	Dry biomass of other seagrass species

measurementTypeID	measurementValue	measurementUnit	measurementUnitID
http://vocab.nerc.ac.uk/collection/P01/current/TEMPPP01/	29.23	Degrees Celsius	http://vocab.nerc.ac.uk/collection/P06/current/UPAA/
http://vocab.nerc.ac.uk/collection/P01/current/SSALSL01/	36	Parts per thousand	http://vocab.nerc.ac.uk/collection/P06/current/UPPT/
http://vocab.nerc.ac.uk/collection/P01/current/DOXYSE02/	6.58	Milligrams per litre	http://vocab.nerc.ac.uk/collection/P06/current/UMGL/
http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL02/	128	Number per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UPMS/
http://vocab.nerc.ac.uk/collection/P01/current/OBSMAXLX/	18	Centimetres	http://vocab.nerc.ac.uk/collection/P06/current/ULCM/
http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.32055	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.05575	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.1469	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.07625	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.0385	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0.02725	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/
http://vocab.nerc.ac.uk/collection/S06/current/S0600087/	0	Grams per square metre	http://vocab.nerc.ac.uk/collection/P06/current/UGMS/

4.8.0.12 Zooplankton biomass & 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.

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	eventRemarks	eventDate	modified
TripNR2547		cruise	2013-07-22T06:58:00+00:00/2013-07-22T16:58:00+00:00	2021-06-23 14:54:00
TripNR2547TripStationNR9781	TripNR2547	stationVisit	2013-07-22T07:13:00+00:00/2013-07-22T07:26:00+00:00	2021-06-23 14:54:00
TripNR2547TripStationNR9781MidasTripActionID23024	TripNR2547TripStationNR9781	sample	2013-07-22T07:22:00+00:00	2021-06-23 14:54:00
TripNR2547TripStationNR9781MidasTripActionID23024TA23024	TripNR2547TripStationNR9781	sample	2013-07-22T07:22:00+00:00	2021-06-23 14:54:00

datasetID	datasetName	locationID	waterBody	country
https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...		Belgian Part of the North Sea	Belgium
https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...	130		
https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...	130		
https://marineinfo.org/id/dataset/4687	LifeWatch observatory data: zooplankton observations...	130		

minimumDepthInMeters	maximumDepthInMeters	decimalLatitude	decimalLongitude	geodeticDatum	footprintSRS
				EPSG:4326	EPSG:4326
0	13.4	51.27083333	2.905	EPSG:4326	EPSG:4326
0	0	51.2687318	2.901797	EPSG:4326	EPSG:4326
3	3	51.2687318	2.901797	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.

eventID	occurrenceID
TripNR2547TripStationNR9781MidasTripActionID23024	TripNR2547TripStationNR9781MidasTripActionID23024occurenceIDTA23024_Amphipoda_sub2_130
TripNR2547TripStationNR9781MidasTripActionID23024	TripNR2547TripStationNR9781MidasTripActionID23024occurenceIDTA23024_Annelida_sub2_130
TripNR2547TripStationNR9781MidasTripActionID23024	TripNR2547TripStationNR9781MidasTripActionID23024occurenceIDTA23024_Anomura_sub2_130
TripNR2547TripStationNR9781MidasTripActionID23024	TripNR2547TripStationNR9781MidasTripActionID23024occurenceIDTA23024_Appendicularia_sub2_130

modified	basisOfRecord	occurrenceStatus	scientificNameID	scientificName
2021-06-22	Occurrence	absent	urn:lsid:marinespecies.org:taxname:1135	Amphipoda
2021-06-22	Occurrence	present	urn:lsid:marinespecies.org:taxname:882	Annelida
2021-06-22	Occurrence	absent	urn:lsid:marinespecies.org:taxname:106671	Anomura
2021-06-22	Occurrence	absent	urn:lsid:marinespecies.org:taxname:146421	Appendicularia

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
TripNR3256TripStationNR17157MidasTripActionID106326		Sampling instrument name
TripNR3256TripStationNR17157MidasTripActionID106326		Sampling net mesh size
TripNR3529TripStationNR19242MidasTripActionID109631UW		Conductivity of the water body
TripNR3529TripStationNR19243MidasTripActionID109634		Concentration of chlorophyll-a per unit volume of the water body
TripNR2547TripStationNR9781MidasTripActionID23024	TripNR2547TripStationNR9781MidasTripActionID23024occurrenceHDTA23024_Annelida_sub2_130	Lifestage
TripNR2547TripStationNR9781MidasTripActionID23024	TripNR2547TripStationNR9781MidasTripActionID23024occurrenceHDTA23024_Annelida_sub2_130	Abundance (ITIS: 64837; WoRMS 882) per unit volume of the water body by image analysis

measurementTypeID	measurementValue	measurementValueID	measurementUnit
http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Planktonnet WP2	http://vocab.nerc.ac.uk/collection/L22/current/TOOL0979/	
http://vocab.nerc.ac.uk/collection/Q01/current/Q0100015/	200		micrometer
http://vocab.nerc.ac.uk/collection/P01/current/CNDCZZ01/	4.05		Siemens per metre
http://vocab.nerc.ac.uk/collection/P01/current/CPHLHPP1/	1.42		Micrograms per litre
http://vocab.nerc.ac.uk/collection/P01/current/LSTAGE01/	unspecified	http://vocab.nerc.ac.uk/collection/S11/current/S1152/	
http://vocab.nerc.ac.uk/collection/P01/current/SDBIOL01/	0.50		specimens/m³

measurementUnitID	measurementDeterminedBy	measurementMethod
http://vocab.nerc.ac.uk/collection/P06/current/UMIC/		
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
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)
	Flanders Marine Institute	identified and counted by image analysis and normalised to a unit volume of water body, validated by human
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

4.8.1 DNA dervived data

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

4.8.1.1 eDNA & DNA derived data

The following example use cases draw on both the GBIF 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.

4.8.1.1.1 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

4.8.1.1.2 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

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: 0.99

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
CCTACGGGNGGCWGCAGGACTACHVGGGTATCTAATCC	341		805R	TCGAGAATTTTTCACAATG...
CCTACGGGNGGCWGCAGGACTACHVGGGTATCTAATCC	341		805R	TCGAGAATTTTTCACAATG...
CCTACGGGNGGCWGCAGGACTACHVGGGTATCTAATCC	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

4.8.2 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, 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.

For datasets with imaging or acoustic data, we strongly recommend including the following terms in your Occurrence table:

- `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 that identified taxa)

Example Resources: Martin-Cabrera et al. (2022) have created a best practices document for plankton imaging data that you can also 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), with 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.

4.8.3 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) 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.

4.8.4 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](https://github.com/tdwg/dwc-for-biologging) 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

5.0.1 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:

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:

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.

5.1 Data quality flags

OBIS performs a number of quality checks on the data it receives. Records may be rejected if the quality does not meet certain expectations. In other cases quality flags are attached to the occurrence records. The checks we perform as well as the associated flags are documented here.

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:

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
```

5.1.1 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.

5.1.1.1 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).

5.1.1.2 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.

5.2 Common Quality Control issues

5.2.1 Missing required fields

5.2.2 Uncertain temporal range

5.2.3 Uncertain geolocation

5.2.4 Uncertain taxaonomic information

5.2.5 Uncertain measurements

5.2.6 Non-marine species

Publishing Data

Chapter 6

Data publication and sharing

OBIS nodes can accept any data files from its data sources or data providers, and they publish these data on their OBIS nodes IPT, which are harvested by central OBIS. The Integrated Publishing Toolkit (IPT) is developed and maintained by the Global Biodiversity Information Facility (GBIF). GBIF maintains an IPT manual. See here for specific OBIS instructions:

6.1 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 OBIS metadata standards and best practices, or check the IPT manual for detailed instructions about the metadata editor.

6.2 Licenses

6.3 IPT Administration

Admin resposibilities

6.4 IPT

Contents

- Introduction
- Installation

- Registration
- Publish your data

- Upload data

- Map to Darwin Core
- Add metadata
- Publish your data
- Publish your data as a dataset paper

6.4.0.1 Introduction

The biodiversity datasets and its metadata are published in OBIS using the Integrated Publishing Toolkit (IPT), developed by GBIF. The IPT software assists the user in mapping data to valid Darwin Core terms and archiving and compressing the Darwin Core content with: (i) 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, and (ii) 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. Al these components (i.e. core file, extension files, descriptor file and metadata file), compressed together (as a .zip file), comprise the Darwin Core Archive.

6.4.0.2 Installation

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.

6.4.0.3 Registration

When you have installed your IPT, please provide the IPT instance URL to the OBIS secretariat, 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 need 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.

6.4.0.4 Publish your data

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 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.

:exclamation: Make sure you have obtained the rights from the data owners to publish their data!

6.4.0.4.1 Create your resource on the IPT

The Integrated Publishing Toolkit (IPT), developed by GBIF, is an open source web application that can be customized by the OBIS node manager. The IPT-instance is used to publish and register all the datasets. To be able to create and manage your own dataset (called a "resource" by GBIF), you will need a user account. Contact your node manager to create one for you.

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. The first thing that needs to be completed is the shortname of your resource. This shortname uniquely identifies your resource (=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
- ...

:exclamation: 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.

You can also create an entirely new resource by uploading an existing archived resource. See the IPT manual section Upload a DwC-A 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 the File upload section 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.

6.4.0.5 Upload data

Uploading your source file to the IPT is easy: go to > your resource overview page > Source Data and click on Choose File. 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.
Once your source file has been uploaded correctly, a source file detail page will 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.

6.4.0.6 Map your data to Darwin Core

Biodiversity data are published in the Darwin Core standard. It 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, and 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. Contact your node manager or the OBIS secretariat at info@iobis.org to guide you through the steps, review your mapping, suggest terms etc.

You can find more information regarding Darwin Core mapping in the IPT manual (including core types, extensions, auto-mapping, default values, value translation, etc.).

6.4.0.7 Publish your data

Go to your resource overview page > Published Release and click Publish. The IPT will now generate your data as Darwin Core, and combine the data with the metadata and package it as a standardized zip-file called a “Darwin Core Archive”. See the IPT manual for more details.

:exclamation: 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 > Public, and it will only be harvested by GBIF when you can Visibility > Registered. The last 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.

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 > Public (see the IPT manual) to make it available to everyone. Warning: please do not do this with your test dataset.

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).

6.4.0.8 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.

6.5 Maintaining published data

6.5.1 Adding a DOI to datasets

6.5.2 User tracking?

6.6 Simultaneous publishing to GBIF

6.6.1 Differences between OBIS and GBIF publication processes

6.7 Update your data in OBIS

Access Data from OBIS

Chapter 7

Data access

7.1 Mapper

- https://mapper.obis.org

The mapper allows users to visualize and inspect subsets of OBIS data. A variety of filters (taxonomic, geographic, time, data quality) is available and multiple layers can be combined in a single view. Layers can be downloaded as CSV files.

7.2 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 including a function reference and a getting started vignette is available at https://iobis.github.io/robis/.

7.3 API

- https://api.obis.org/

Both the mapper and the R package are based on the OBIS API which can be used by third party developers as well.

7.4 Full exports

- https://obis.org/data/access/

Full exports of the quality controlled presence records as CSV or Parquet (see below).

7.5 Finding your own data

7.6 Citing Data from OBIS

General OBIS citation:

```
OBIS (YEAR) Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. www.obis.org.
```

Use the following format to cite data 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, you can, in addition to cite 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

7.7 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

7.8 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.

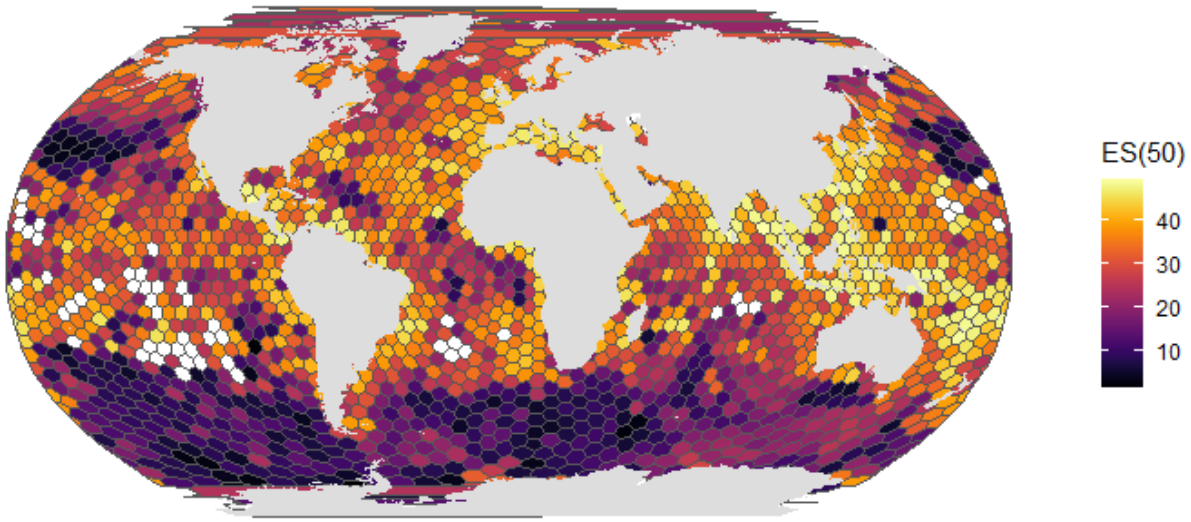


Figure 7.1: screenshot

Additional Resources

Chapter 8

Other Resources

In this section we highlight resources created by collaborators.

8.1 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.

8.2 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.

8.3 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.