OBIS Manual

Table of contents

Making better use of OBIS data

OBIS is not just a repository for biodiversity data. It's also a community of practice that want to leverage the use of ocean biodiversity data. Here you will find tutorials and tools to help you make better use of data from OBIS (and other resources).

You can contribute with tutorials following the guidelines available at our repository.

Don't know where to start? We suggest reading the OBIS manual where you will find the most complete guide about how OBIS works and the different types of data available.

Check also the available packages and pipelines for R and other languages, and our tool to find the right Darwin Core schema for your data.

Part I

Overview

1 Introduction

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

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

1.1 Guidelines on the sharing and use of data in OBIS

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

1.2 Acknowledgements

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

1.3 Data Policy

1.3.1 Guidelines on the sharing and use of data in OBIS

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

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

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

Unless data are collected through activities funded by IOC/IODE, neither UNESCO, IOC, IODE, the OBIS Secretariat, nor its employees or contractors, own the data in OBIS and they take no responsibility for the quality of data or products based on OBIS, or the use or misuse that people may make of them nor can it control or limit the use of any data or products accessible through its website, other than through the use of a published Data Sharing and Use Terms and Conditions.

1.3.1.1 Data sharing agreement

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

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 and GithHub (see below) so you are more likely to receive a quick answer to your question by posting in either place, as the Secretariat receives many requests.

You can submit issues and questions on relevant Github repositories:

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

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

Part II

Contributing data to OBIS

2 What can you contribute and how?

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

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

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

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

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

1. First, you must identify which OBIS node is best suited to host your published data. If you would like to publish to GBIF at the same time, that is also possible. If your organization is already affiliated with a GBIF node with which you must publish from, OBIS can also harvest from GBIF nodes.
2. Second, you must determine the structure of your data and which format will best suit your dataset. OBIS follows Darwin Core Archive (DwC-A) standards for datasets, and currently follows a star schema format. This format is based on relational databases. If you are unfamiliar with such database structures, or would like to refamiliarize yourself with them, please read here
3. Then, you need to actually format your data according to OBIS and DwC-A standards and guidelines
4. Once formatted, you should run a series of quality control measures to ensure you are not missing any required information and that all standards are being met. This helps ensure all data published in OBIS is formatted in a standardized way. When published in OBIS, OBIS provides a quality report to inform data owners and users of any quality control issues. By completing quality control before you publish your dataset you ensure there are fewer errors to fix later.
5. Now that your dataset is ready for publishing, the relevant metadata must be filled in, and then published on the previously identified IPT.

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

2.1 Why publish data to OBIS

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

- **F - Findable**

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

- **A - Accessible**

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

- **I - Interoperable**

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

- **R - Reusable**

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

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

- Your dataset can be associated with a DOI, allowing for your dataset to be more easily cited. By ensuring your dataset citation is complete you will ensure you are being cited properly.
- Publishing your dataset with OBIS makes it easier to set it up as a Data paper, which generates value for you and other researchers.
- There are social benefits to data publishing as your work becomes integrated into a wider dataset. It gives both you and your data more visibility. This can lead to more opportunities for collaboration and further career development as a researcher or professional.
- Your data can be incorporated into larger analyses to better understand global ocean biodiversity, helping to shape regional and international policies.

2.2 How to handle sensitive data

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

- Location data on endangered or protected species
- Information regarding a commonly poached species

- Species or locations that have an economic impact (positive or negative)

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

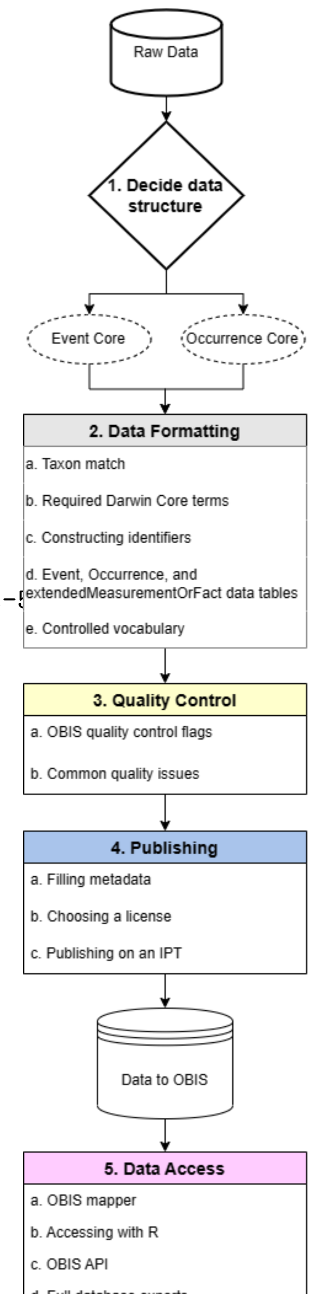
- Generalizing location information by: Obtaining regional coordinates using MarineRegions, Getty Thesaurus of Geographic Names, or Google Maps
- Using the OBIS Map tool to generate a polygon area with a Well-Known Text (WKT) representation of the geometry to paste into the `footprintWKT` field.
- Delay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

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

2.3 OBIS Data Life Cycle

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

1. Data structure



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

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

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

Depending on your data structure, you will then format data into a DwC-A format with the appropriate Core table (Event or Occurrence) with any applicable extension tables. Any biotic or abiotic measurements will be moved into the extendedMeasurementOrFact table. Before proceeding to the publishing stage, there are a number of quality control steps to complete.

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

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

2.3.0.1 Who is responsible for each phase?

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

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

2.4 Biodiversity data standards

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

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

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

We also provide some dataset examples for your reference.

2.4.1 Darwin Core

Contents

2.4.1.1 Introduction to Darwin Core

Darwin Core is a body of standards (i.e., identifiers, labels, definitions) that facilitate sharing biodiversity informatics. It provides stable terms and vocabularies related to biological objects/data and their collection. Darwin Core is maintained by TDWG (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases). Stable terms and vocabularies are important for ensuring the datasets in OBIS have consistently interpretable fields. By following Darwin Core standards, both data providers and users can be certain of the definition and quality of data.

2.4.1.1.1 History of Darwin Core and OBIS

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

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

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

2.4.1.2 Darwin Core (DwC) terms

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

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

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

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

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

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

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

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

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

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

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

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

- decimalLatitude
- decimalLongitude
- coordinateUncertaintyInMeters
- geodeticDatum
- footprintWKT

- minimumDepthInMeters
- maximumDepthInMeters
- minimumDistanceAboveSurfaceInMeters
- maximumDistanceAboveSurfaceInMeters
- locality
- waterBody
- islandGroup
- island
- country
- locationAccordingTo
- locationRemarks
- locationID

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

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

The following DwC terms are related to the Class *Material-Sample*:

- materialSampleID

2.4.1.3 Darwin Core guidelines

2.4.1.3.1 Taxonomy and identification

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

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

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

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

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

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

Examples:

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

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

Data from Benthic fauna around Franz Josef Land.

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

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

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

Examples:

scientificName	scientificNameAuthorship	scientificNameID	taxonRank	identificationQualifier	taxonConceptID
Pelagia	Péron & Lesueur, 1810	urn:lsid:marinespecies.org:taxname:135262	genus	gen. nov.	Pelagia gen. nov.
Pelagia benovici	Piraino, Aglieri, Scorrano & Boero, 2014	urn:lsid:marinespecies.org:taxname:851656	species	sp. nov	Pelagia benovici sp. nov
Gadus	Linnaeus, 1758	urn:lsid:marinespecies.org:taxname:125732	genus	cf. morhua	Gadus cf. morhua
Polycera	Cuvier, 1816	urn:lsid:marinespecies.org:taxname:138369	genus	cf. hedgpethi	Polycera cf. hedgpethi
Tubifex	Lamarck, 1816	urn:lsid:marinespecies.org:taxname:137392	genus	?	Tubifex tubifex(Müller, 1774)?
Tubifex	Lamarck, 1816	urn:lsid:marinespecies.org:taxname:137392	genus	sp. inc.	Tubifex tubifex(Müller, 1774)sp. inc.
Brisinga	Asbjørnsen, 1856	urn:lsid:marinespecies.org:taxname:123210	genus	gen. inc.	Brisinga gen. inc.
Uroptychus compressus	Baba & Wicksten, 2019	urn:lsid:marinespecies.org:taxname:1332465	genus	sp. inc.	Uroptychus compressus sp. inc.
Eurythenes	S. I. Smith in Scudder, 1882	urn:lsid:marinespecies.org:taxname:101607	genus	sp. DISCOLL.PAP.JC165.674	Eurythenes sp.DISCOLL.PAP.JC165.674

scientificName	scientificNameAuthorship	scientificNameID	taxonRank	identificationQualifier	taxonConceptID
Paroriza	Hérouard, 1902	urn:lsid:marinespecies.org:taxname:123467	genus	sp.[unique123]aff.pallens	Paroriza sp.[unique123]aff. pallens
Aristeidae	Wood-Mason in Wood-Mason & Alcock, 1891	urn:lsid:marinespecies.org:taxname:106725	family	stet.	Aristeidae stet.
Nematocarcinus	Milne-Edwards, 1881	urn:lsid:marinespecies.org:taxname:107015	genus	sp.indet.	Nematocarcinus sp.indet.
Brisinga	Asbjørnsen, 1856	urn:lsid:marinespecies.org:taxname:123210	genus	gen.inc.	Brisinga gen.inc.
Brisinga costata	Verrill, 1884	urn:lsid:marinespecies.org:taxname:17825	species	sp.inc.	Brisinga costata sp.inc.

2.4.1.3.2 Occurrence

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

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

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

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

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

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

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

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

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

Example:

collectionCode	occurrenceID	catalogNumber	occurrenceStatus
SluiceDock_benthic_1976/1981	SluiceDock_benthic_1976_1	SluiceDock_benthic_1976_1	present
SluiceDock_benthic_1976/1981	SluiceDock_benthic_1976_2	SluiceDock_benthic_1976_2	present

SluiceDock_benthic_1976/1981	SluiceDock_benthic_1979-07/1980-06_1	SluiceDock_benthic_1979-07/1980-06_1	present

Data from [A summary of benthic studies in the sluice dock of Ostend during 1976-1981](#).

2.4.1.3.3 Record level terms

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

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

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

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

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

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

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

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

2.4.1.3.4 Location

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

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

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

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

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

Some examples of WKT strings:

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

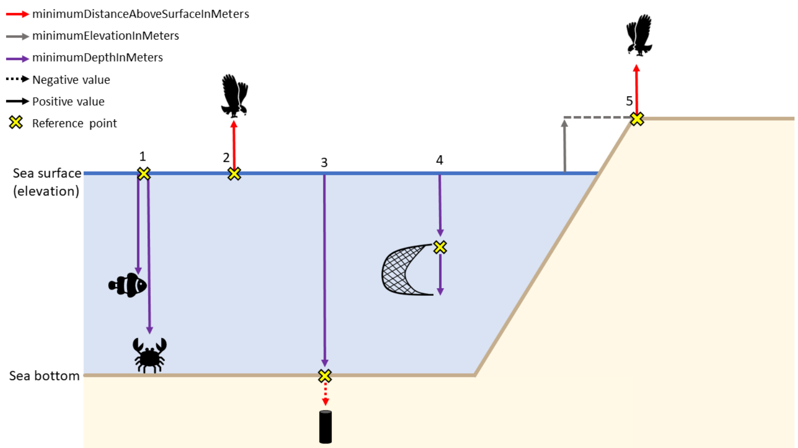
Example:

decimalLatitude	decimalLongitude	geodeticDatum	coordinateUncertaintyInMeters	footprintWKT	footprintSRS
38.698	20.95	EPSG:4326	75033.17	LINESTRING (20.31 39.15, 21.58 38.24)	EPSG:4326
42.72	15.228	EPSG:4326	154338.87	LINESTRING (16.64 41.80, 13.82 43.64)	EPSG:4326
39.292	20.364	EPSG:4326	122083.27	LINESTRING (19.05 40.34, 21.68 38.25)	EPSG:4326

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

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

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



Depth scenario examples:

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

2.4.1.3.5 Event

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

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

2.4.1.3.6 Time

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

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

2.4.1.3.7 Sampling

Information on `sampleSizeValue` and `sampleSizeUnit` is very important when an organism quantity is specified. However, with OBIS-ENV-DATA it was felt that the extended MeasurementorFact (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

- OBIS holds more than just species occurrences: the ENV-DATA approach
  - ExtendedMeasurementOrFact Extension (eMoF)
  - eDNA & DNA derived data Extension
  - A special case: habitat types
- 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. *(Note you must refine search to Technical Reports from 2021 to identify Duncan et al.'s report)*

2.4.3 Relational databases: the underlying framework of OBIS

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

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

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

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

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

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

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

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

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

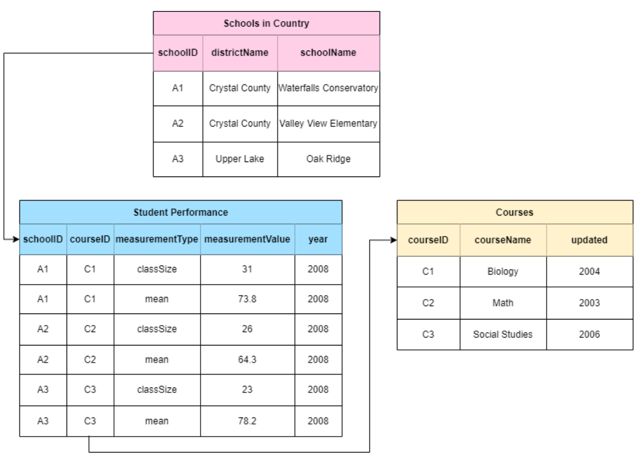


Figure 2.1: *An example of how a relational database works. Three tables show the (1) student performance (blue table) in (2) different schools (pink table) in a fictional country, and (3) the names of the courses (yellow table). Information between each table is linked by the use of identifiers, indicated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

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

2.4.3.1 How to avoid redundancy

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

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

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

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

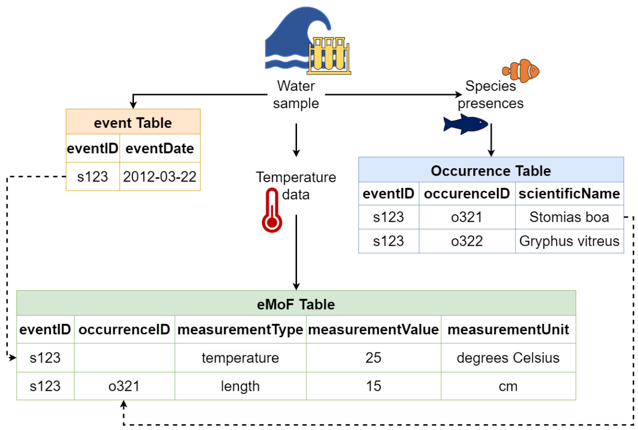


Figure 2.2: *Example of how the sample data is distributed to Core and Extension tables, and how these tables are connected in OBIS*

combination with extension files can really simplify and reduce the number of times data are repeated.

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

2.4.4 Ecological Metadata Language

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

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

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

- `title [xml:lang="..."]`: A good descriptive `title` is indispensable and can provide the user with valuable information, making the discovery of data easier. Multiple titles may be provided, particularly when trying to express the title in more than one language (use the "xml:lang" attribute to indicate the language if not English/en).

- `creator` ; `metadataProvider` ; `associatedParty` ; `contact` : These are the people and organizations responsible for the dataset resource, either as the creator, the metadata provider, contact person or any other association. The following details can be provided:

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

- `pubDate`: The date that the resource was published. Use ISO 8601.