A short guide to publishing data in EurOBIS and EMODnet Biology

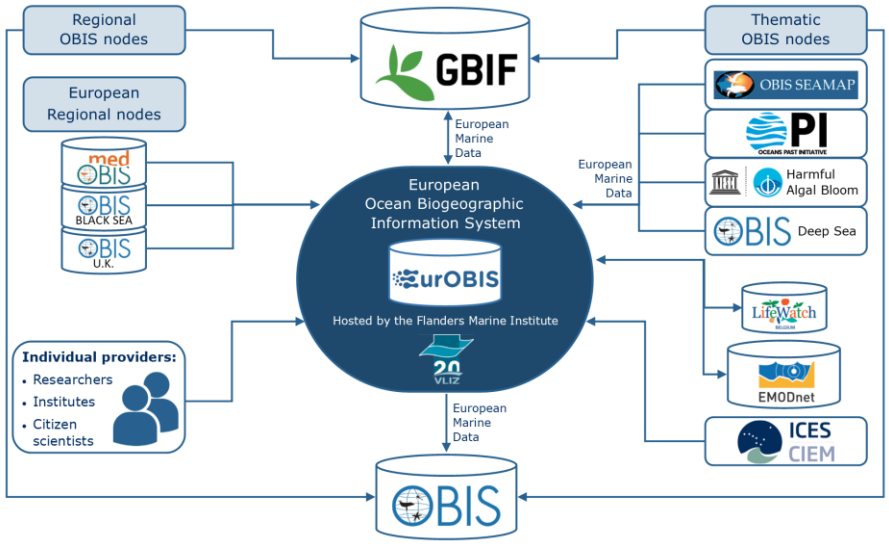
VLIZ

EurOBIS
European Ocean Biogeographic Information System

EMODnet
European Marine Observation and Data Network

Introduction

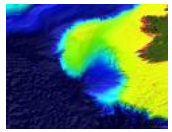
The aim of this document is to provide guidance on how to make data available through EurOBIS. There are several ways to share your biodiversity data and make them available in EurOBIS and EMODnet Biology. Whatever path is chosen, both data and metadata need to be formatted following accepted standards to allow for the efficient discovery of datasets and data interoperability. Data in the EurOBIS database are synchronized with EMODnet and will flow to OBIS and GBIF.



EMODnet and EMODnet Biology

The European Marine Observation and Data Network (EMODnet) is a network of organisations that work together to observe the sea, process the data according to international standards and make that information freely available as interoperable data layers and data products.

EMODnet provides access to European marine data across seven discipline-based themes:


Bathymetry


Biology


Chemistry


Geology


Human Activities


Physics


Seabed Habitats

EMODnet Biology aims to provide a single access point to European Marine Biodiversity Data and Products by assembling individual datasets from various sources and processing them into interoperable data products for assessing the environmental state of ecosystems and sea basins. It is built upon the World Register of Marine Species (WoRMS) and the European Ocean Biogeographic Information System (EurOBIS).

Specific objectives of EMODnet Biology:

- Provide public access to search, download and viewing tools for data, metadata and data products of marine species occurring in European marine waters;
- Create specific biological data products to illustrate the temporal and geographic variability of occurrences and abundances of marine phytoplankton, zooplankton, macro-algae, angiosperms, fish, reptiles, benthos, birds and sea mammal species with a priority to develop those required for support management, policy, planning and education;
- Improve harmonisation of differing methodologies and strategies for data management under common protocols, data formats and quality control procedures (by adopting EMODnet and INSPIRE standards);

- Ensure consistent distribution of data by making use of relevant open web services for various user applications;
- Provide tools for spatial, temporal and taxonomic queries.

The EMODnet Biology data portal provides free access to data on temporal and spatial distribution of marine species and related measurements from all European regional seas.

What type of data can you submit?

Geographical scope	European and non-European, provided they are collected by European institutes. Water column and seabed data, including coastal zones
Temporal scope	All
Data types	Raw data Data products Metadata (when dataset is restricted)
Parameters	Taxon name, geographical location and observation date Observation: presence only, abundance, and biomass. Related parameters: including biotic (e.g. length measurements, etc) and abiotic (e.g. sediment composition, water temperature, salinity, etc)
Data origin	Small scale, one-off scientific research data Large scale and/or long-term monitoring Museum collections and literature data
Taxonomic scope	All different groups of marine species (EMODnet Biology themes: algae, angiosperms, benthos, birds, fish, mammals, phytoplankton, reptiles and zooplankton)
Data accessibility	CC licence required

Where can you submit your data?

1. EurOBIS online submission form

The EurOBIS portal provides a submission form similar to the one found through the EMODnet Biology portal. The dataset will be handled by VLIZ in collaboration with the data provider and will be made available in EurOBIS if desired.

2. EMODnet Biology online submission form

The EMODnet Biology portal provides a submission form for users who wish to contribute their biodiversity data. The dataset will be handled by VLIZ in collaboration with the data provider and will be made available in EMODnet Biology and EurOBIS if desired.

3. EMODnet Data Ingestion portal

EMODnet Ingestion web portal was created to facilitate submission of marine data of any discipline. An online submission form and a help service assist the user filling in the metadata and submit their data in the most appropriate format. The user will be put in contact with the corresponding National Oceanographic Data Center (NODC) in charge of processing the data and, for biology data, implementing the pathway from Ingestion to the EurOBIS and EMODnet Biology portals.

4. Integrated Publishing Toolkit

The GBIF Integrated Publishing Toolkit (IPT) is a freely available open source web application that makes it easy to share biodiversity-related information. The user is guided by the application through a series of steps to format their data, fill in the necessary metadata and publish the dataset. If you have a good number of datasets or you will need to publish or update datasets regularly, you might consider setting up your own IPT by following these guidelines.

Requirements

The publication of datasets in EurOBIS or EMODnet Biology adheres to a number of requirements:

- The dataset needs to be described: a minimum set of **metadata** has to be provided
- The data have to be organized and formatted according to the Darwin Core **standard** (DwC)
- The data will have to undergo essential **quality checking**
- VLIZ recommends assigning a **Digital Object Identifier** (DOI) to the dataset

All of the steps above will be performed by an online tool or facilitated by an institution (e.g. NODC) that will handle your dataset with your collaboration. Each step will be succinctly described:

1. Metadata: describing your data

Metadata will help other users to better understand the content of the data, it will extend data longevity and facilitate data discovery and reuse.

There are many advantages in providing good metadata:

- It is a way of organising electronic resources and it is necessary to make sure your dataset can be found (discoverability).
- It is essential for interoperability: it enables understanding of the data by humans and machines
- It is needed to indicate the origin of the dataset and who to contact in case of any questions.
- It makes your dataset easy to use and understand.
- It helps potential users to decide whether the dataset is useful for their purposes or not without the need to download it first.
- It informs potential users on data restrictions
- It informs on how potential users are to cite the dataset upon use in academic papers and other publications

In EMODnet Biology we keep a catalogue (http://www.emodnet-biology.eu/data-catalog?module=dataset&show=search) with metadata records for all the datasets known to exist in order to increase their discoverability. The metadata records clearly indicate and provide a link if a dataset is available through the portal.

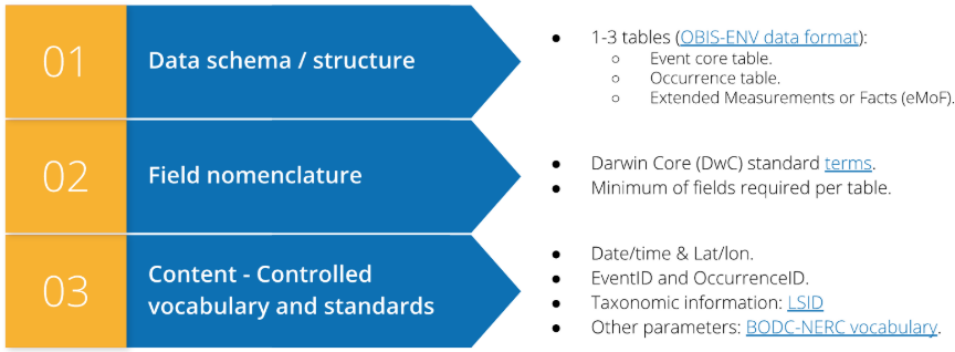
Through the submission forms the mandatory and optional fields are indicated and the most important ones, together with some standard advice on how to fill them are listed below:

A. Full name of the dataset: Should be descriptive, meaningful and concise (e.g. following a pattern of What, Where, When, How, Why), allowing users to easily know what to find in the dataset
B. Abstract: should be written in a way that helps potential users to understand if the data may be of their interest. Should be a short description indicating data type or origin, taxa or functional groups included, when and where data were collected
C. Person filling in this form: a person or organization that can be contacted in case there are questions related to the dataset: name and institute of the data provider, an email address
D. Citation: Should contain the author(s) of the dataset, year of publication, dataset title, publisher and identifier (if available). It can also mention the version and resource type. It is the equivalent of a publication reference
E. Keyword(s): different keywords that are especially important and/or descriptive for the dataset. You can either pick them from the standard ASFA list, fill them in by yourself or combine keywords from the list and keywords entered by your own
F. Data license: clarifying under which conditions the dataset can be used. We request the use of Creative Commons licenses
G. Temporal cover: for the dataset in YYYY-MM-DD format (or other ISO 8601 compliant formats)
H. Geographical cover: List area(s) or location(s) where the data were collected. We advise using Marine Regions to find the adequate geo-units
I. Taxonomic cover: Overview of the taxa included in the dataset. We recommend using the World Register of Marine Species for internationally accepted taxa

2. Data format

Standardization facilitates data discovery, integration, sharing and interoperability. Both EurOBIS (and OBIS) and EMODnet Biology use the OBIS-ENV-Data (https://obis.org/manual/dataformat/#obis-env-data), based on the Darwin Core Archive (DwC-A https://dwc.tdwg.org/terms/#theterms) standard for biodiversity, used by GBIF. The DwC is a body of standards with a list of defined terms that allow your data to be understood and used by everyone. This standard determines the way your data will be structured (i.e. number of tables), the number, the name and the content of the fields for each of these tables. EMODnet Biology and OBIS also use the BODC vocabularies (https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/) to standardise parameters that are not covered by DwC.

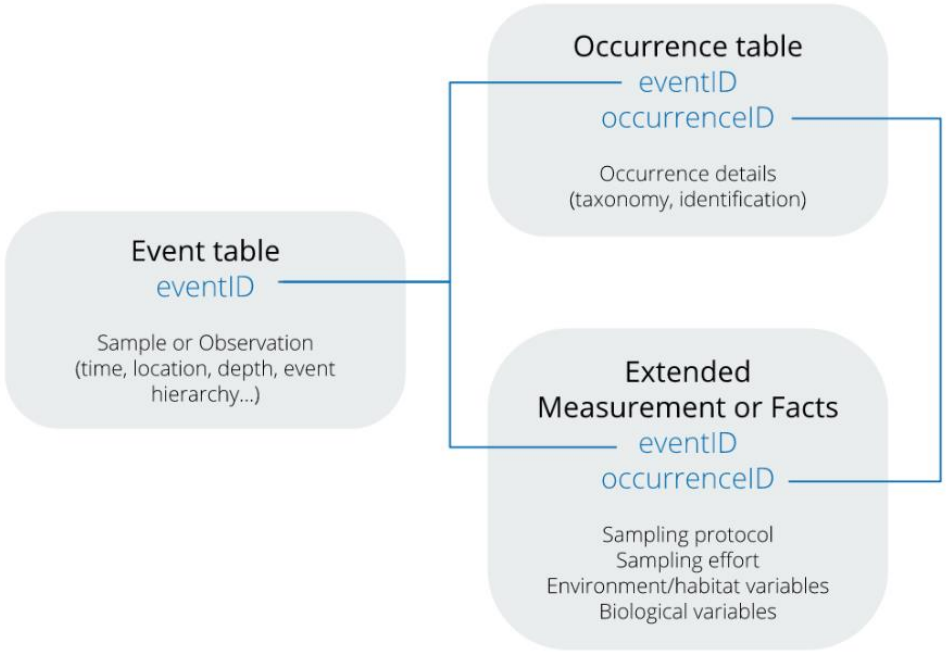
We can divide the data format in three main blocks, each of them with their own specific standards:



2.1 Data structure

Marine biological data often include measurements related to habitat features, such as physical and chemical variables of the environment, and biotic measurements (such as body size, counts, abundance and biomass, etc) as well as details regarding the nature of the sampling or observation methods, equipment, and sampling effort.
In order to capture all this information, the conceptual data model of the Darwin Core Archive is a "star schema" with a core table in the center of the star (the sampling event) and extension tables radiating out of the center.

In practice, EMODnet Biology and OBIS use a subset of 1 to 3 tables to represent the data but in most cases, we will use the three tables. The three tables are related via the eventID and the occurrenceID



2.2 Field nomenclature

The field names of each of the 3 tables have to follow the Darwin Core terminology http://rs.tdwg.org/dwc/terms/. This spreadsheet template contains a detailed summary of the most common fields for each table.

2.3 Content

Data interoperability is achieved through the use of controlled vocabularies, besides the field names, the content or the data itself has to follow certain data standards. An explanation on how the mandatory fields are populated is available in the OBIS manual.

For example:

- Date related fields have to be ISO 8601 compliant.
- Latitude and longitude must be in decimal degrees and referenced to the WGS84 datum
- ScientificNameID must contain a LSID (created by WoRMS for each taxon) that can be found using the WoRMS taxon match tool (or the LifeWatch WoRMS webservice if too many taxons are to be matched).
- Other parameters: The BODC vocabularies are essential for the Extended Measurements or Facts (eMoF) table, the fields where they should be used are measurementTypeID, measurementValueID and measurementUnitID. The most often used ones within OBIS-ENV format are identified below:
  - Q01- OBIS sampling instruments and methods attributes
  - P01- BODC Parameter Usage Vocabulary
  - S11- Biological entity life stage terms
  - S10- BODC parameter semantic model biological entity gender terms
  - L05- SeaDataNet device categories
  - L22- SeaVoX Device Catalogue
  - C17- ICES Platform Codes
  - P06- Approved data storage units

It is important to note that:

- Upon submitting your data via EurOBIS, EMODnet Biology or EMODnet Ingestion, you will be assisted by VLIZ or the correspondent NODC respectively.
- If you decide to upload the data using the IPT you should use the Darwin Core mapping tool, which allows a match between the fields in the source file with the appropriate DwC terms
- Additional information such as biometrics of the observed taxa (e.g. body length) or even non-biological parameters (e.g. water temperature, sediment grain-size, etc) are supported by the Darwin Core schema and (Eur)OBIS (see De Pooter et al., 2017), and therefore welcome

3. Data quality checking

Before any data are made available through EurOBIS or EMODnet Biology, they go through a series of quality control procedures, which aim to improve data quality and completeness and add value, thus improving the overall EurOBIS and EMODnet databases. The checks include:

- All the necessary metadata and data fields are filled in
- All taxon names present are matched to their standard name in the World Register of Marine Species (WoRMS)
- All supplied coordinates are given in WGS84 and have possible values (-90 to 90, -180 to 180), and all supplied dates are in ISO 8601 format
- The correct units are given for abundance and/or biomass (if these parameters are supplied)
- Check for issues in eMoF table for the BODC Vocabulary

The above QC tests are performed in a semi-automatic way on the data once they have been uploaded to the database. If problems are encountered, the data provider will be contacted by either the corresponding NODC or VLIZ to clarify the issues.
The data providers are encouraged to prepare their datasets and QC their data using the online verification tools developed by VLIZ, like the LifeWatch data validation and QC services.
The LifeWatch portal offers useful web services that can help you validate your data. We refer to the manual for more details on how to:

- Plot the coordinates on your file with the Show on map tool.
- Check if the coordinates values are possible (global values, in the marine environment) with the Check OBIS file tool
- Check if the date values provided are correct with the Check OBIS file tool
- Check if there are any mismatches in the eMoF table

4. Digital Object Identifier (DOI)

A Digital Object Identifier (or DOI) is a character string used to uniquely identify an object. Metadata describing the object is stored in association with the DOI name, including a URL which leads to where the object can be found.

Associating DOIs to scientific publications has not only **increased the traceability** of the cited literature but also simplified the maintenance of citation indexes which serve today to assign academic credit to scientists for their work. As is true for claims based on information from other publications, in scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited and more and more journals are requesting data DOIs prior to publishing. Additionally, there is growing international support for the idea that **dataset citations should also lead to academic credit**. VLIZ can aid you in the process of assigning a DOI to your dataset. For more information, you can read the VLIZ DOI guidelines or contact data@vliz.be.

Resources

The current guidance is not to be used as a stand-alone document and we advise the following resources to be used alongside:

- De Pooter et al, 2017, Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences
- Ocean Teacher Global Academy (OTGA) courses (UNESCO-led project):
- 2016, OBIS-INDEEP training & workshop
- 2015, Marine Biogeographic Data Management (contributing and using OBIS)
- Through the OTGA there is also a specific training section about OBIS.