What is EMODnet Biology

You will learn a bit more about the EMODnet project and the Biology portal. More specifically, you will have a look at how the data flows to EMODnet Biology and its relation with other data systems with a focus on (marine) biodiversity.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	What is EMODnet Biology
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:18 PM

Table of contents

EMODnet: the European Marine Observation and Data Network

EMODnet Biology: unlocking European Marine Biodiversity Data

EMODnet Biology data policy

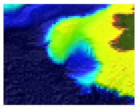
EMODnet Biology data flow

EMODnet: the European Marine Observation and Data Network

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.

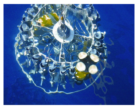
EMODnet provides access to European marine data across seven discipline-based themes:

Bathymetry



Biology

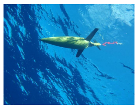
Chemistry



Geology

Human Activities

Physics



Seabed Habitats

More about EMODnet: http://www.emodnet.eu/

EMODnet Biology: unlocking European Marine Biodiversity Data

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 and the European Ocean Biodiversity Information System.

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 algae, benthos, birds, fish, mammals, phytoplankton, reptiles and zooplankton 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 INSPIRE standards);
- Ensure consistent distribution of data by making use of relevant open webservices for various user applications;
- Provide tools for spatial, temporal and taxonomic queries.

More about EMODnet Biology: http://www.emodnet-biology.eu/

EMODnet Biology data policy

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

The consortia partners should make data available using one following 3 Creative Commons licenses.

- CC-0
- CC-BY
- CC-BY-NC

Learn more: https://creativecommons.org/licenses/

EMODnet Biology data flow

You have read that EMODnet Biology is built upon the European Ocean Biodiversity Information System (EurOBIS), but **what does that mean?**

The following presentation explains how data flows towards EMODnet Biology and how it relates to other biodiversity data systems, such as the Ocean Biodiversity Information System.

FAIR Principles in EMODnet Biology

Brief overview of FAIR principles and how they apply to EMODnet Biology

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	FAIR Principles in EMODnet Biology
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:24 PM

Table of contents

FAIR Principles

DOIs and ORCIDs

FAIR Principles

The FAIR Guiding Principles for scientific data management and stewardship were first published in 2016 in the *Scientific Data* journal. They put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.

The principles apply to both data and metadata and are:

1. Findable
2. Accessible
3. Interoperable
4. Reusable

This presentation briefly introduces these principles but more information can be found in Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

DOIs and ORCIDs

EMODnet Biology encourages the use of Digital Object Identifiers (DOI) to datasets. This is however only done when thy are requested by the data providers. A DOI is a static, permanent link to the data, allowing it to be citable, and more visible. It also provides evidence for data claims and the citations can contribute to scholarly credit for the data creators. More detailed information can be found in the book "Data harvest and DOIs" under the EMODnet Biology data harvesting section of this course.

Within EMODnet Biology we also encourage our data providers to use Open Researcher and Contributor Identifier (ORCID) as it is persistent, unique digital identifier used to disambiguate researchers from one another.

You might have noticed that an increasing number of journals and publishing companies are requesting DOIs and ORCIDs when submitting documents for publication.

EMODnet Biology data management

You will have a quick look at the data management cycle of EMODnet Biology, including an introduction to the data format (data structure and standards used).

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	EMODnet Biology data management
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:25 PM

Table of contents

Overview of EMODnet Biology data management process

Dataset description: metadata

Data processing: formatting the data according to EMODnet Biology (OBIS) standards (I)

Data processing: formatting the data according to EMODnet Biology (OBIS) standards (II)

Data processing: structure of the data

Data processing: field nomenclature

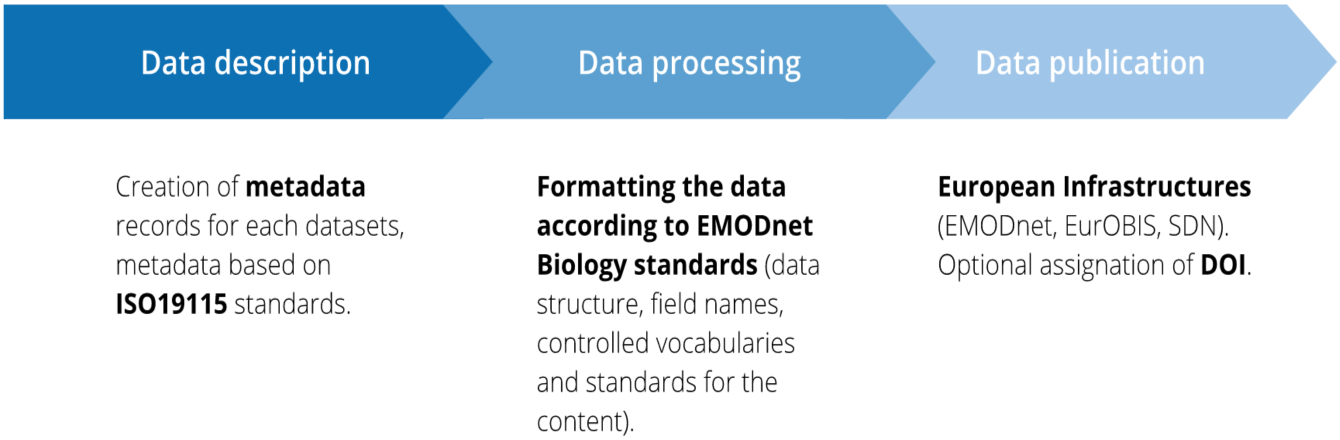
Data processing: content standards

Data publication

Overview of EMODnet Biology data management process

EMODnet Biology's data management is composed by three main steps:

- Data description
- Data processing
- Data publication



Data description

Creation of **metadata** records for each datasets, metadata based on **ISO19115** standards.

Data processing

Formatting the data according to EMODnet Biology standards (data structure, field names, controlled vocabularies and standards for the content).

Data publication

European Infrastructures (EMODnet, EurOBIS, SDN). Optional assignation of **DOI**.

Dataset description: metadata

There are many advantages in providing good metadata:

- It is necessary to make sure your dataset can be found (discoverability)
- It is needed to indicate the origin of the dataset and who to contact
- It makes your dataset easy to use and understand
- It helps potential users to decide whether the dataset is useful 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 publications

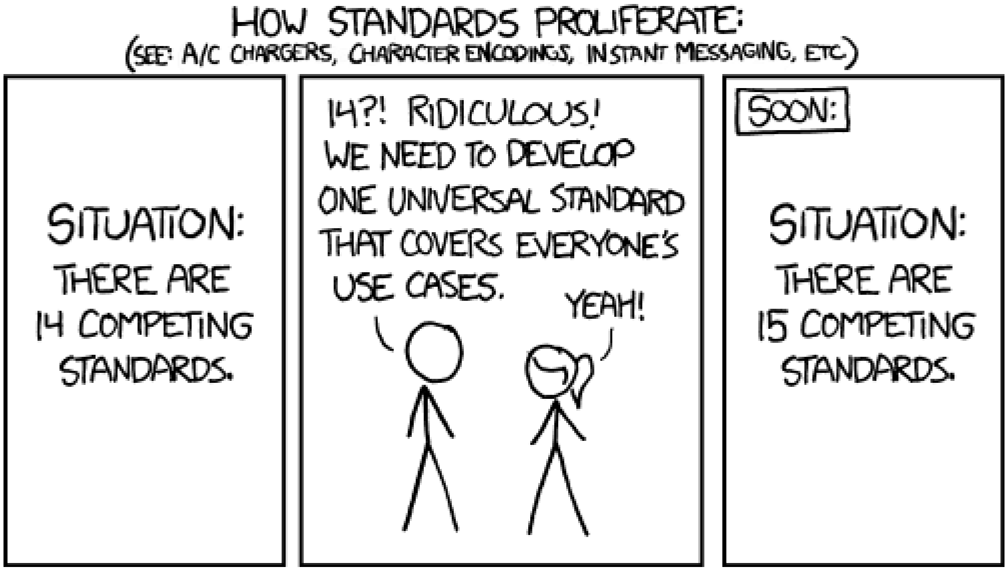
In EMODnet Biology we keep a catalogue with metadata records for all the datasets, including those available through the portal and those not yet available.

Later on in the course you will learn more about how to provide good quality metadata.

Data processing: formatting the data according to EMODnet Biology (OBIS) standards (I)

Both EMODnet Biology and OBIS (and GBIF) rely on Darwin Core Archive (DwC-A), the standard for publishing biodiversity data. This standard determines the way the data will be structured (i.e. number of tables we need to provide), the number, name and the content of the fields for each of these tables.

EMODnet Biology and OBIS, unlike GBIF, also use the BODC vocabularies to standardise parameters not covered by DwC. These vocabularies are part of one of DwC's extensions, the DwC OBIS-ENV format.

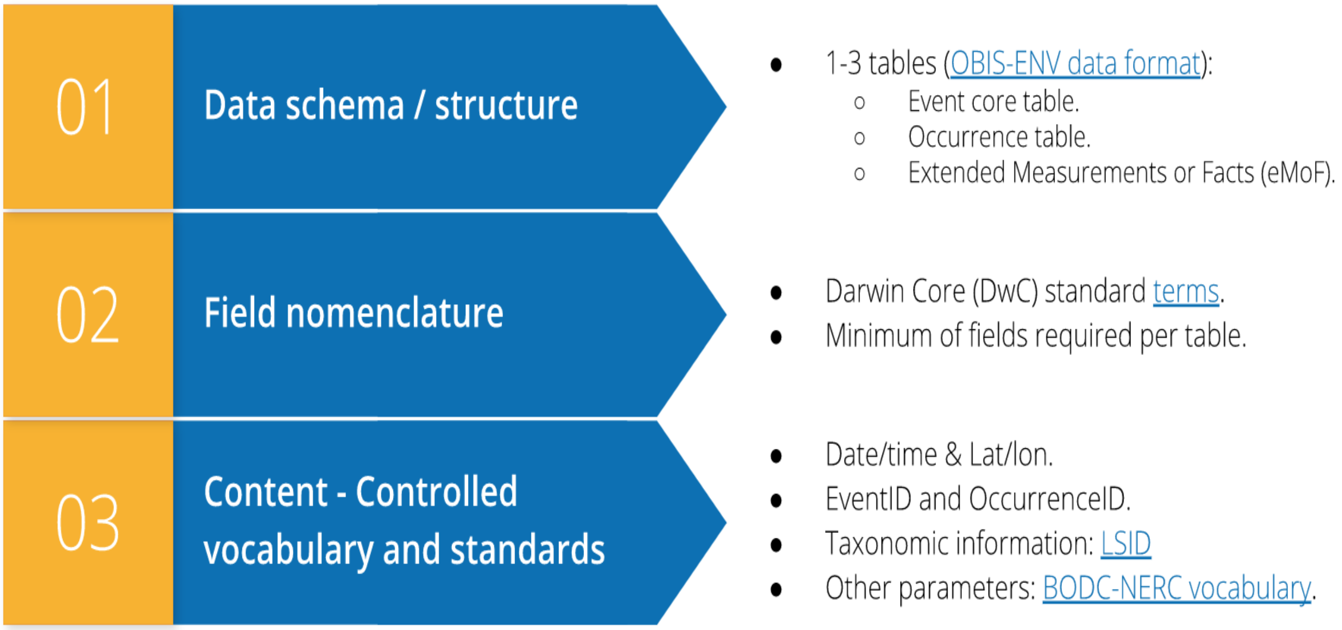

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

(by xkcd)

Data processing: formatting the data according to EMODnet Biology (OBIS) standards (II)

We can think of the data processing as three main blocks, each of them with their own specific characteristics:

- **Data structure**: the conceptual data model of the Darwin Core Archive is a "star schema" with a core table in the center of the star 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. In most cases, we will use the three tables.
  - Event (core) table: to store sample and/or observation information (time, location, depth, event hierarchy).
  - Occurrence table: to store occurrence details (taxonomy, identification, organismID).
  - Extended Measurements or Facts (eMoF) table: to store sampling information and additional biological and/or abiotic measurements.
- **Field nomenclature**: the field names of each of the 3 tables have to follow the Darwin Core terminology. There is a minimum number of fields required per table.
- **Content**: besides the field names, the content, or the data itself, has to follow certain standards. For example, the date-related fields have to be ISO 8601 compliant.



01 Data schema / structure

- 1-3 tables (OBIS-ENV data format):
  - Event core table.
  - Occurrence table.
  - Extended Measurements or Facts (eMoF).

02 Field nomenclature

- Darwin Core (DwC) standard terms.
- Minimum of fields required per table.

03 Content - Controlled vocabulary and standards

- Date/time & Lat/lon.
- EventID and OccurrenceID.
- Taxonomic information: LSID
- Other parameters: BODC-NERC vocabulary.

Later in the course each of the these blocks will be explained in detail with exercises.

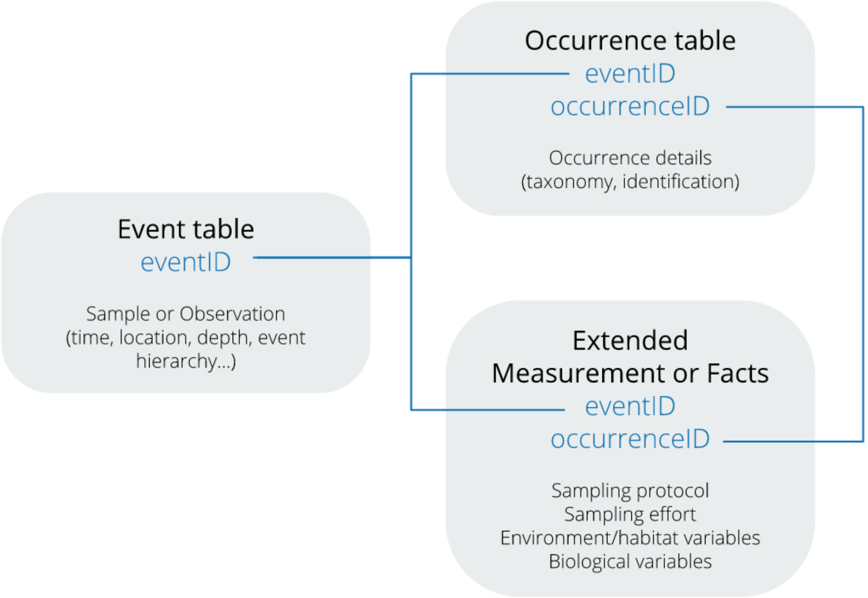
Data processing: structure of the data

Marine biological data often includes 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, EMODnet Biology and OBIS make use of three tables of the DwC star schema.

In this case, the sampling event (Event table) is the central entity, which is linked to two extensions: the Occurrence and eMoF (extended Measurements or Facts) tables. This schema is often referred to as OBIS Event Core or OBIS-ENV.

The three tables are related via the eventID and the occurrenceID.

- The eMoF table is used in combination with the Event Core and the Occurrence table to capture sampling information, abiotic and biotic measurements
- The occurrenceID is used to link biotic measurements in the eMoF table with the the Occurrence table
- The eventID links the eMoF table to the Event Core



Data processing: field nomenclature

The field names of each of the 3 tables have to follow the Darwin Core terminology. There are many DwC terms but not all of them are required for EMODnet Biology. In the "Data processing" book the section "Most relevant DwC terms" provides detailed information on the terms used in each table.

Data processing: content standards

As was mentioned before, the content or the data itself have to follow certain standards and use certain vocabularies. Here are some examples:

- Date-related fields have to be ISO 8601 compliant: e.g. 2018-03-02 [YYYY-MM-DD]
- Latitude and longitude have to be in decimal degrees and referenced to the WGS84 datum

An overview of the required format for the content of the different fields is available here.

Later on in the course you will learn how to populate the fields that require a specific format:

- EventID and OccurrenceID.
- Taxonomic information: LSID
    - *Gadus ruber; Gadus callarias; etc. standardized to: Gadus morhua (urn:lsid:marinespecies.org:taxname:322691)*
- Other parameters (eMoF table): BODC-NERC controlled vocabulary.
    - Biomass per g dry weight; dry weight biomass; biomass (g dry weight); etc... can be standardized to: Dry weight biomass (in assayed sample) of biological entity specified elsewhere (http://vocab.nerc.ac.uk/collection/P01/current/ODRYBM01/)

Data publication

Once your dataset fully complies with the EMODnet Biology standards, the next step will be to publish it. If you remember the EMODnet Biology data flow, the dataset will be first published in EurOBIS, from where it will flow to EMODnet Biology, LifeWatch, OBIS and, finally GBIF.

There are different options available to get your data published in EMODnet Biology. The preferred option is by setting an IPT instance (Integrated Publishing Toolkit), an open source software tool that is used to publish and share biodiversity datasets through the GBIF network. There is a specific module on the course about publishing your data with IPT.

DOIs

Another aspect you might consider is to assign a DOI to your dataset. To meet the growing demand towards dataset **traceability** and **citability**, VLIZ collaborates with **DataCite** to provide the scientific community the opportunity to formally publish their datasets by assigning **Digital Object Identifiers** (DOIs).

Learn more about why and how to assign a DOI to your dataset here.

Metadata in EMODnet Biology

You will learn about EMODnet Biology Catalogue records and the procedure to create an maintain them. You will learn which are the mandatory and which optional fields in the catalogue and which information they contain.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Metadata in EMODnet Biology
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:26 PM

Table of contents

Introduction to metadata

The EMODnet Biology Catalogue

Best metadata practices (I)

Metadata best practices (II)

Dataset citation

Metadata exchange format - IPT

Introduction to metadata

Metadata is often defined as data about the data. It means describing your datasets in a structured way so that your data can be found and understood. EMODnet Biology considers receiving good metadata essential for the following reasons:

- Metadata facilitates data reuse and sharing: it ensures data are more easily interpreted, analysed and processed by the data originator and others. Missing relevant metadata hinders the use of the data. It also enables that datasets that were designed for a single purpose to be reused for other purposes
- Metadata is a way of organising electronic resources, making them discoverable
- Metadata is essential for interoperability: it enables understanding of the data by humans and machines

In EMODnet Biology we keep a catalogue with metadata records for all the datasets available through the portal and some that are not yet available.

The EMODnet Biology Catalogue

All EMODnet Biology datasets are described in the EMODnet Biology Catalogue. The catalogue is part of a larger metadata system, the Integrated Marine Information System (IMIS). As IMIS is hosted by Flanders Marine Institute (VLIZ), it is focused on Flanders and supplemented by the scientific output of projects involving VLIZ. It contains metadata about all people, institutes, publications, projects and datasets that are about or involved in marine science and links these different modules together.

This means that if you want to submit a dataset to EMODnet Biology, the following metadata records will need to be created in the EMODnet Catalogue:

- A **dataset** metadata record: A record containing all the metadata for your dataset. When the data have been made available this record will include a link to where the data can be downloaded from. For example: http://www.emodnet-biology.eu/data-catalog?module=dataset&dasid=4355
- Metadata records for all **people** who contributed to the dataset and should be acknowledged in the metadata record. For example: http://www.emodnet-biology.eu/data-catalog?module=person&persid=7528
- Metadata records for all **institutes** who contributed to the dataset and should be acknowledged in the metadata record. For example: http://www.emodnet-biology.eu/data-catalog?module=institute&insid=36
- Metadata records for all **publications** based on or describing the dataset. For example: http://www.emodnet-biology.eu/data-catalog?module=ref&refid=243062

Please note that all these records are interlinked, which allows for easy discovery of the scientific output of an institute or a data provider.

Creating and editing EMODnet catalogue records

The content of the EMODnet catalogue is managed and edited solely by the EMODnet Data management team. To create a dataset record in the EMODnet catalogue, we require you to fill in this excel template file and email it to bio@emodnet-biology.eu.

If you find an error in your metadata record, you can report these to us directly through the [report an error in this record] button at the top of each metatada record, or via email to bio@emodnet-biology.eu.

Best metadata practices (I)

Good metadata should help the user to find the data and understand what the data are, thus we encourage that it is as complete as possible. Here is an overview of what information is mandatory, highly recommended and optional for the EMODnet Biology metadata:

Person providing the metadata: name, institute, email	**Mandatory**
Dataset title in English	**Mandatory**
Dataset title in original language (and language)	Optional
Contact person for the dataset: name, institute, email	**Mandatory**
Data creator(s): (name), institute	**Mandatory**
Other person(s) associated with the dataset	Highly recommended
Dataset citation	**Mandatory**
License or terms of use	**Mandatory**
Abstract	**Mandatory**
Extended description	Highly recommended
Geographical coverage	**Mandatory**
Temporal coverage	**Mandatory**
Taxonomic coverage	**Mandatory**
Themes	**Mandatory**
Keywords	Optional
Websites	Optional
Publications related to the dataset	Highly recommended

The next page provides an explanation on what is expected in the mandatory fields. This information is also available in the metadata template. You can also find extended information in the VLIZ Data Submission Guidelines and in the guide to publish biodiversity data in EurOBIS and EMODnet Biology.

Metadata best practices (II)

Dataset title

Choose a title that is descriptive, meaningful and concise. A good title allows a user to a first assessment whether the dataset is useful for the intended purpose. EMODnet Biology recommends including the region and time period of sampling. For example: "Benthic data from the Southern Irish Sea from 1989-1991"/"Zoobenthos of the Kyklades (Aegean Sea) from a survey in 2009".

If a dataset has already been published with a title that does not conform these above requirements, the original title should be kept, as re-publishing the dataset with a completely new title is likely to cause confusion and maybe data duplication.

Abstract

Provide a good abstract that can help potential users understand if the data are of interest to them. This should be a short description providing an indication of the content of the dataset (e.g. *Distribution of dinoflagellates causing harmful algal bloom in the Mediterranean sea collected from literature. The data contains only HABs that appeared prior to 2004.*).

The extensive description, if included should be written in plain language and include the answers to the following questions: WHAT, WHERE, WHEN, HOW, WHY and WHO.

Dataset contact

The resource contact, person or organisation that can be contacted in case a user has questions related to the dataset or seeks collaboration: name and institute of the data provider, an email address.

Data creator(s)

They are either the persons or institutes responsible for the creation/maintenance of the dataset.The data creators will be mentioned in the citation.

Person providing the metadata

The person who created the content of the metadata record. This is usually the person who gave a title the dataset and wrote the abstract.

Keywords

Add a few keywords that enhance dataset discovery, they can originate from the ASFA list or can be created by you.

Data licence

Specify the usage of the dataset or a data licence: i.e.: under which conditions the dataset can be used. EMODnet Biology requests the use of Creative Commons licences:

- CC-0 Public Domain Dedication - https://creativecommons.org/publicdomain/zero/1.0/
- CC-By Attribution - https://creativecommons.org/licenses/by/4.0/
- CC-By-NC Attribution-Non Commercial - https://creativecommons.org/licenses/by-nc/4.0/

Temporal scope

Give the date or the temporal scope covered by your dataset. If there is a large temporal gap (+2 years) this should be indicated.

Geographical scope

Describe the spatial or geographical extent of the data by listing the area(s) or location(s) where data was collected (e.g. South Atlantic Ocean, Belgian part of the North Sea). We encourage the use of Marine Regions to find adequate geo-units.

Taxonomic scope

Provide an overview of the taxonomic scope present in the dataset (e.g Crustacea, Pisces). Make sure to pick the appropriate taxonomic level. We recommend the use of the World Register of Marine Species to find the adequate taxa.

Themes

Specific for EMODnet Biology, it should describe the themes covered by the dataset: Algae, Angiosperms, Benthos, Birds, Fish, Mammals, Phytoplankton, Reptiles, Zooplankton.

Publications related to the dataset

Here you can list all publications which are (partly) based on data from this dataset, that describe it (e.g. a data-paper) or that were used to compile the data from it (if the dataset is based on literature studies).

Dataset citation

The dataset citation is one of the most important metadata aspects, as it is the equivalent of a publication reference. The proposed dataset citation **should reference the dataset itself**, and not a paper using or describing the dataset.

Core elements of the dataset citation:

Ryan, William B.F. (2009), Global Multi-Resolution Topography (GMRT) synthesis. Version 0.3. Integrated Earth Data Applications (IEDA). http://dx.doi.org/10.1594/IEDA/100001 Accessed 17 November 2017.

- Data creator(s) -> the people and/or organisations responsible for the intellectual work to develop the dataset.
- Publication Year -> when the particular version of the dataset was first made available for use (and potential citation) by others.
- Title -> the formal dataset title
- Version -> the precise version of the data used, if available. Careful version tracking is critical to accurate citation.
- Archive and/or Distributor (~Publisher) -> the organization distributing or caring for the data, ideally over the long term.
- Locator/Identifier -> this could be a URL but ideally it should be a persistent service, such as a DOI, Handle or ARK, that resolves to the current location of the data in question.
- Access Date and Time -> because data can be dynamic and changeable in ways that are not always reflected in release dates and versions, it is important to indicate when online data were accessed.

For more information check the following guide:

Ball, A. & Duke, M. (2015). ‘How to Cite Datasets and Link to Publications’. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

Dataset citation in EMODnet Biology

We require that the citation of your dataset contains the Data Creator(s), Publication Year and Title. It should also contain a DOI, when available. If applicable it can also mention the publisher and the version.

- *Belmonte, G. (2010) Zooplankton - Crustacea from the Taranto Seas. University of Salento, Department of Biological and Environmental Science and Technologies, Laboratory of Zoology and Marine Biology (LZMB), Italy. http://dx.doi.org/10.14284/57*
- *Bio-environmental research group; Institute of Agricultural and Fisheries research (ILVO), Belgium; (2016): Macrobenthos monitoring at long-term monitoring stations in the Belgian part of the North Sea between 1979 and 1999. http://dx.doi.org/10.14284/201*

Metadata exchange format - IPT

EMODnet Biology has its own metadata catalogue and we require you to provide us with metadata for it using a template, however if you submit your data through the IPT, you will have to fill in metadata for the IPT record too.

There is no automated system to transfer the metadata from the IPT to the EMODnet Biology catalogue. Whilst harvesting your data from the IPT, EMODnet Biology may do a brief assessment on whether the metadata on your IPT and in our catalogue are the same.

In order to ensure that the metadata in the EMODnet Biology Catalogue is correct, we require you to advise us when a correction in the catalogue is needed.

You will learn more about the IPT and the metadata to be provided in the "How to publish the dataset through IPT?" book in this course.

Technical metadata and data entity integrity

You will learn about the importance of technical metadata and what to keep in mind upon receiving the dataset you are meant to process for EMODnet Biology.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Technical metadata and data entity integrity
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:26 PM

Table of contents

Technical metadata and data entity integrity

Technical metadata and data entity integrity

Apart from data which are typically stored in a separate database, metadata also comprises a more technical aspect which makes sure the data can be used. Checks related to the technical metadata include the following aspects:

- **Parameters**: Is it unambiguously clear, for all parameters, what was actually measured (e.g. wet weight vs ash free dry weight)?
- **Units**: Are the associated units recorded for all parameters**?**
- **Instruments, sampling descriptors and protocols.** Are they provided and if so is it unambiguously clear what is described?
- **Dates (and time):** Is the time zone specified (UTC, GMT, MET, CST, …)?
- **Coordinates:** Is the datum and the coordinate system recorded? How were the coordinates recorded and what is the error associated to them?
- **Relational integrity**: if the data are provided in database format, do all ID's resolve?
- **Duplication** (and 'contradictions'): are the same specimens listed only a single time for each event? Do parameters have only a single value for each occurrence or event?

These technical metadata are relevant as EMODnet Biology is not only interested in biogeographical data, but also in all biotic or abiotic measurements collected together with the data as they add ecological context to the observation. In order for EMODnet Biology to integrate data from different sources it is vital that parameters are unambiguously described, otherwise it is impossible to know if parameters from different datasets describe the same measurements.

Of vital importance, for comparing different data from different surveys or creating biogeographical data products, are the sampling descriptor data. For example for benthos data 3 terms, at least, are crucial to be able to create a presence/absence product: the sampling instrument, sampling surface area and the sieve mesh size.

- **Sampling instruments** differ in the way they collect sediments. In modern benthology, box corers are preferred over grabs as they enable samples in which all sediment depths are equally represented. However, unlike grabs, box corers produce turbulence at the sample's sediment surface so some smaller epibenthic species can be swept away.

- The final benthic organism abundance data are generally expressed per surface area (e.g. number of individuals or number of species per m²). However, it is very relevant to know the original sample size, because environmental conditions in sediments are very heterogeneous and organisms tend to display aggregative distributions. This means that a slight increase in **sampling surface area** can have dramatic effects on the sampled abundance.

- After being extracted, organisms are selected through a sieve with a specific mesh size. Traditionally, the "macrobenthos" is selected with a **mesh size** of 1 mm, but sometimes other sizes are used, this is especially the case for old surveys.

In the following sections you will encounter some examples of issues regarding the technical metadata and data entity integrity.

Field nomenclature

This topic focuses on Darwin Core terms. You will have a look at the most relevant terms used in EMODnet Biology and will learn which terms are mandatory to be used in each of the three tables (Event, Occurrence and Extended Measurements or Facts). Using a demo exercise, you will practice the mapping of existing column names to the DwC terms.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Field nomenclature
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:26 PM

Table of contents

Data processing

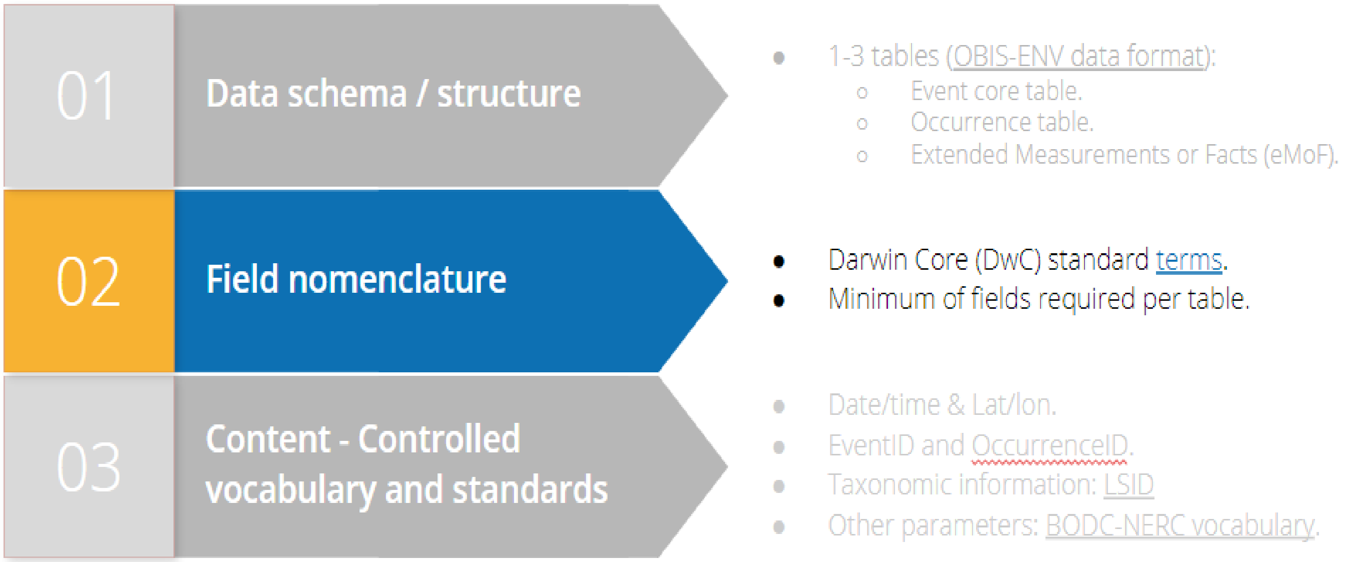
What are Darwin Core (DwC) terms?

Most relevant DwC terms

Data processing

As you have seen in the introduction, you can think of data processing as three main blocks, **data structure**, **field nomenclature**, and **content**, each of them with their own specific standards.

This book will focus on **Field nomenclature**. You start with this topic because, by applying Darwin Core terminology to your column names, you can later assign each of your columns to the correct table more easily.



01 Data schema / structure

- 1-3 tables (OBIS-ENV data format):
  - Event core table.
  - Occurrence table.
  - Extended Measurements or Facts (eMoF).

02 Field nomenclature

- Darwin Core (DwC) standard terms.
- Minimum of fields required per table.

03 Content - Controlled vocabulary and standards

- Date/time & Lat/lon.
- EventID and OccurrenceID.
- Taxonomic information: LSID
- Other parameters: BODC-NERC vocabulary.

What are Darwin Core (DwC) terms?

As you've seen in the book Contributing data sets to EMODnet Biology, EMODnet Biology relies on Darwin Core as a data formatting standard.

Darwin Core is a body of standards for biodiversity informatics. It provides stable terms and vocabularies for sharing biodiversity data. Darwin Core is maintained by TDWG (Biodiversity Information Standards, formerly The International Working Group on Taxonomic Databases).

Darwin Core terms correspond to the column names of your data set. The field names of each of the three DwC tables have to follow the DwC terminology. There are many DwC terms, but not all of them are necessary.

Most relevant DwC terms

A list of all possible DwC terms can be found in the Darwin Core Reference Guide. The DwC terms that are most relevant to EMODnet Biology (and OBIS) format are the following (those in bold are mandatory):

Event table

eventID, **datasetName**, **institutionCode**, **eventDate**, **decimalLatitude**, **decimalLongitude**, parentEventID, type, minimumDepthInMeters, maximumDepthInMeters, coordinateUncertaintyInMeters, footprintWKT, modified

Occurrence table

eventID, **occurrenceID**, **scientificName**, **scientificNameID**, **occurrenceStatus**, **basisOfRecord**, collectionCode, catalogNumber, scientificNameAuthorship, kingdom, taxonRank, identificationQualifier, modified

Extended Measurement or Fact (eMoF) table

eventID, **measurementType**, **measurementTypeID**, **measurementValue**, measurementID, occurrenceID, measurementValueID, ***measurementUnit*** *(mandatory for measurements)*, measurementUnitID, measurementAccuracy, measurementRemarks.

More information about each of the other terms can be found in the book Data standardisation - events and occurrences and Data standardisation - eMoF.

To be able to correctly complete the assignment and quiz, it is necessary to understand the what all these terms stand for. **Make sure to check out them in the Darwin Core Reference Guide.**

Data standardisation - events and occurrences

This topic gives detailed information about how to standardise the content of your data set. You will learn how to format and populate important fields of the data set in a standardized matter.
A big part of this chapter covers the formatting of ID fields. They form the important connections between the different tables and prevent redundancy in data storage.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Data standardisation - events and occurrences
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:27 PM

Table of contents

Common terms - Introduction

Common terms - occurrenceID, institutionCode, and collectionCode

Common terms - eventID, and parentEventID and type
Formatting a basic eventID
Hierarchy and type

Common terms - eventDate

Common terms - basisOfRecord, occurrenceStatus

Taxonomy - Introduction to WoRMS and the link with EMODnet Biology

Taxonomy - WoRMS Taxon Match Tool: preparing an input file

Taxonomy - WoRMS Taxon Match Tool: obtaining the LSIDs

Taxonomy - WoRMS Taxon Match Tool: the incredible return

Taxonomy - Which information should be included in the dataset for EMODnet?

Taxonomy - What if a taxon did not match with WoRMS?

Locations - Exact positions

Locations - Start and end positions - Linestrings

Locations - General areas and regions - Polygons

Common terms - Introduction

For all Darwin Core terms there are guidelines towards what kind of content they can contain. For some terms there are also guidelines to how this content should be structured. Many of these guidelines are determined by the Darwin Core standard itself, but in some cases OBIS and EMODnet Biology have more restrictive ones than those determined by Darwin Core.

The first chapter of this book discusses the EMODnet Biology guidelines for following relevant terms: occurrenceID, institutionCode, collectionCode, eventID, eventDate, basisOfRecord, and occurrenceStatus.

Further chapters in this book will pay special attention to the taxonomic terms (scientificName, scientificNameID, scientificNameAuthor, identificationQualifier), and the location related terms (decimalLatitude, decimalLongitude, coordinatePrecisionInMeters and footprintWKT).

Some fields require a separate controlled vocabulary and non-DwC terms to be standardised. These are terms like Lifestage, Sex, sampling descriptors, and data set specific measurements like sediment grain size, and will be discussed in the book Data standardisation - eMoF.

Common terms - occurrenceID, institutionCode, and collectionCode

If the dataset does not yet contain (globally unique) occurrenceIDs, then they should be created. The institutionCode, collectionCode and datasetName should be added too. All of these fields are very useful for making sure that the right data file is associated to the right data set. This makes controlling for duplicates easier.

The **occurrenceID** is an identifier for an occurrence record and should be persistent and globally unique.
There are no guidelines yet on designing occurrenceIDs, but in the absence of a persistent globally unique identifier, an ID can be constructed by combining the institutionCode, the collectionCode and the catalogNumber (or autonumber in the absence of a catalogNumber).

The **institutionCode** is the code of the legal entity for which your data were collected (e.g UGhent, INBO, IEO, IFREMER). If you wish to specify a specific department within that legal entity, you should use both the code for the legal entity and the one for the department (e.g. UGhent - MARBIOL). For data describing museum collections this term refers to the institute which holds the specimen.

The **datasetName** should be the same as the title of your data set, and the **collectionCode** should be an abbreviation of that data set title.
If you already have a globally unique occurrenceID, you can choose to only fill in the **datasetName**.
For data describing museum collections, the collectionCode refers to the name of the collection where the specimen can be found. This code should be unique for the data set, at least within the institute, and preferably also globally unique.

The **catalogNumber** can be used in two different ways depending on the type of data.
For data describing museum collections: this term refers to the label of the specimen, or the jar in which the specimen is kept within a museum collection.
For other types of data: if the data set has identifiers for the occurrence records which are unlikely to be globally unique, they can be filled out here. If the data set has no such identifier, an autoNumber can be assigned and filled out here.

An example of different unique codes, and the occurrenceID which combines these unique codes:

institutionCode	collectionCode	catalogNumber	occurrenceID
UGhent	NSBS	123	UGhent_NSBS_123
UGhent	NSBS	456	UGhent_NSBS_456

Common terms - eventID, and parentEventID and type

The **eventID** is an identifier for a sampling related activity. This can for example refer to a Cruise which was undertaken, the visit of a station, a sample which was collected, a transect which was conducted, a quadrat which was assessed or a subsampling action.

The **parentEventID** is an optional field for the identifier which relates to the eventID for the 'parent' action. If the eventID of the records refers to a subsample, then the parentEventID can be filled out with the eventID that is used to record the sample.

The **type** is used to specify to which type of activity the event refers to. A controlled vocabulary is needed. Terms included are "cruise", "stationVisit", "transect", "quadrat", "sample", "subsample". Others can be added upon need.

Formatting a basic eventID

When we have replicate samples, each different sample needs to have a globally unique eventID and each subsample needs its own globally unique eventID. If eventIDs are already present in the dataset, it is highly recommended to use these eventIDs (in their formulation) as to ensure the traceability to the database. If there are no eventIDs present in your dataset yet, we have to construct them according to following guidelines.

Creating a table with unique values requires that you have some insight of the data. ***What combination of fields makes an event unique?*** Typically, unique combinations of date/time, coordinates and depth will define unique events. An eventID can thus be built using these values and a parentEventID.
Consider the following format:
"{parentEventID} & - & {other value} & - & {ID}".
Here ID refers to an ID which is likely to be unique in your dataset. By including the parent structure in the eventID we increase the likelihood of this ID being globally unique.

Optionally, for complicated hierarchical datasets, it can be helpful if the *type* is included in the eventID too. To ensure global uniqueness, a collectionCode can be added in front of the eventID. The design would then be "{parentEventID} & - & {type} & : & {ID}". Below you can find an example which includes the type in the eventID.

eventID	parentEventID	type
POHJEdatabase-stationVisit:10098		stationVisit
POHJEdatabase-stationVisit:10098-sample:15976	POHJEdatabase-stationVisit:10098	sample
POHJEdatabase-stationVisit:10098-sample:15976-subSample:sieve_05mm	POHJEdatabase-stationVisit:10098-sample:15976	subSample

Hierarchy and type

Imagine replicates of a sample were taken without indicating a new date/time. This should be reflected by the presence of a different EventID and by the hierarchy with parentEventID, as well as the field Type (“sample” vs. “replica”).

eventID	**parentEventID**	**type**	**eventDate**	**decimalLongitude**	**decimalLatitude**
cruise_1		cruise			
cruise_1:station_1	cruise_1	stationVisit		-12.0190	33.9069
cruise_1:station_1:grab_1	cruise_1:station_1	sample	2016-01-02T16:02		
cruise_1:station_1:grab_2	cruise_1:station_1	sample	2016-01-02T16:24		
cruise_1:station_1:grab_1:subsample_1	cruise_1:station_1:grab_1	subsample			

Note that in our example, we have not created parentEventIDs for the stations (given by LocationID). You could do it but, if you look closely, the coordinates are given at the level of the sample. Adding a parentEventID for the stations would be meaningful if all the samples within the same station shared the same coordinates (and/or depth).

The inclusion of the column parentEventID in the Event Core makes it possible to create hierarchies among events. In biological oceanography, for example, data are often gathered using research vessels that visit several stations during a given cruise and deploy different instruments at each station. The event hierarchy allows for the creation of one event record for each cruise (parent event), one event record for each visit to a station (child event), different event records for each sampling activity at a station (grandchild events) and, if applicable, different event records for subsamples (great-grandchild events).

The event hierarchy makes it possible to record differences in sampling time, location and depth while grouping these samples together in the same station visit. In addition, there is the added benefit of keeping all data at the appropriate levels and thus reducing data duplication to a minimum. (From De Pooter et al. 2017)

This means that you will only need to write down the cruise name at the level of the parent event cruise. Similarly, if the coordinates for all the samples within one station are the same, you don’t need to duplicate the coordinates for all the sample (grandchild) events. However, if the coordinates are slightly different for each sample within the same station, this variation can be recorded by adding the coordinates at the level of the sample (grandchild) event. We still know that these samples belong together because it is reflected by the hierarchy: they all correspond to the same station (child) event.

We also need to add a column to specify the event type. This column complements the hierarchy information given by EventID and parentEventID. To fill in the "type" column, controlled vocabulary is recommended. Commonly used terms included are "cruise", "stationVisit", "transect", "quadrat", "sample", "subSample". Others can be added upon need.

Common terms - eventDate

ISO 8601 standard

The date and the time at which of the sampling event or the observation is stored in **eventDate**. All information stored in eventDate should use the **ISO 8601 standard.**

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:

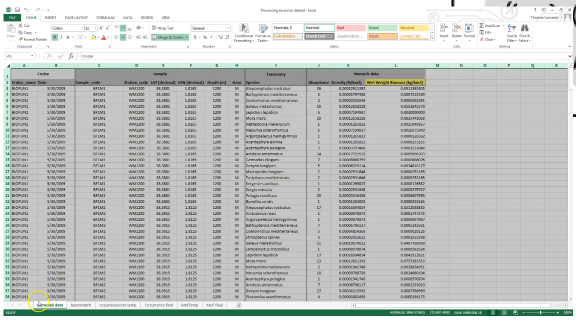
1973-02-28T15:25:00
2005-08-31T12:11+12
1993-01-26T04:39+12/1993-01-26T05:48+12
2008-04-25T09:53
1948-09-13
1993-01/02
1993-01
1993

PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS *THE* CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27⁄2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
(111+1)-1)×3/3-1/3³
/1101 02/27/20/13 01237
HISSss
2-27-13

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). These ordinal dates should NOT be used for EMODnet Biology.

Working with dates in excel

Dates are tricky to format when working with Excel. Typically Excel will recognise the content as a date and will ask you how to display it. However, how the date is displayed in a date field differs from how it's stored or exported. The safest way work with dates is to store them as text. A useful function in excel for converting dates to text is "=TEXT(D2, "YYYY-MM-DD")" or if you have time specified "=TEXT(D2, "YYYY-MM-DDThh:mm")". See the video below for how to use it.


Processing excercise dataset - Excel
Thomas Lanssens
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW
Cut
Copy
Format Painter
Paste
Clipboard
Calibri 11
Font
Wrap Text
Merge & Center
Alignment
General
Number
Conditional Formatting
Format as Table
Normale 2 Normal Bad Good Neutral
Calculation Check Cell Explanatory ... Input Linked Cell
Styles
Insert Delete Format
Cells
AutoSum
Fill
Clear
Sort & Filter
Find & Select
Editing
corrected data taxonmatch occurrencecore temp Occurrence final MofTemp MoF final
READY

A1 fx Cruise

	Cruise		Sample						Taxonomy	Numeric data		
2	Cruise_name	Date	Sample_code	Station_code	LAT (decimal)	LON (decimal)	Depth (m)	Gear	Species	Abundance	Density (N/km2)	Wet Weight Biomass (kg/km2)
3	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Alepocephalus rostratus	26	0.00032911392	0.0911392405
4	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Bathypterois mediterraneus	3	0.00003797468	0.0007215190
5	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Coelorinchus mediterraneus	2	0.00002531646	0.0005063291
6	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Galeus melastomus	10	0.00012658228	0.0531645570
7	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Lepidion lepidion	6	0.00007594937	0.0030000000
8	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Mora moro	10	0.00012658228	0.0835443038
9	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Nettastoma melanurum	1	0.00001265823	0.0015949367
10	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Nezumia sclerorhyncus	6	0.00007594937	0.0016075949
11	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Argyropelecus hemigymnus	1	0.00001265823	0.0000126582
12	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Acantephyra eximia	1	0.00001265823	0.0000253165
13	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Acantephyra pelagica	3	0.00003797468	0.0002531646
14	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Aristeus antennatus	14	0.00017721519	0.0005696203
15	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Gennadas elegans	7	0.00008860759	0.0000886076
16	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Geryon longipes	5	0.00006329114	0.0034810127
17	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Macropodia longipes	2	0.00002531646	0.0000253165
18	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Pasiphaea multidentata	2	0.00002531646	0.0000253165
19	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Sergestes arcticus	1	0.00001265823	0.0000126582
20	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Sergia robusta	2	0.00002531646	0.0000379747
21	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Pelagia noctiluca	20	0.00025316456	0.0058607595
22	BIOFUN1	5/30/2009	BF1M1	WM1200	38.3881	1.8165	1200	M	Bonellia viridis	1	0.00001265823	0.0000253165
23	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Alepocephalus rostratus	17	0.00016504854	0.0512038835
24	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Arctozenus risso	1	0.00000970874	0.0001747573
25	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Argyropelecus hemigymnus	1	0.00000970874	0.0000097087
26	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Bathypterois mediterraneus	7	0.00006796117	0.0005145631
27	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Coelorinchus mediterraneus	5	0.00004854369	0.0009029126
28	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Etmopterus spinax	3	0.00002912621	0.0081553398
29	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Galeus melastomus	11	0.00010679612	0.0407766990
30	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Lampanyctus crocodilus	1	0.00000970874	0.0000582524
31	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Lepidion lepidion	17	0.00016504854	0.0042912621
32	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Mora moro	13	0.00012621359	0.0757281553
33	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Nettastoma melanurum	2	0.00001941748	0.0028834951
34	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Nezumia sclerorhyncus	10	0.00009708738	0.0024660194
35	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Acantephyra pelagica	2	0.00001941748	0.0000970874
36	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Aristeus antennatus	7	0.00006796117	0.0002233010
37	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Geryon longipes	27	0.00026213592	0.0087766990
38	BIOFUN1	5/30/2009	BF1M2	WM1200	38.3915	1.8125	1200	M	Plesionika acanthonotus	4	0.00003883495	0.0000194175

AVERAGE: 5964.375635 COUNT: 4900 SUM: 16992506.18

Common terms - basisOfRecord, occurrenceStatus

basisOfRecord

This term records the "specific nature" of the data record. In EMODnet Biology and OBIS it is used to explain on what type of evidence the claim of the occurrence of the taxon is based. This term is mandatory for IPT, which means that you need to assess the data set metadata and choose the most suited value. The most relevant options for EMODnet Biology are:

- **materialSample**: if the claim is based on sample specimen. Someone held the specimen closely and examined it to determine the taxon it belongs to. It is for specimens that were removed from the environment.
- **preservedSpecimen**: when the sampled specimen is deposited in a collection. If this term is chosen then you should provide the institutionCode, collectionCode and catalogNumber of the specimen. The institutionCode refers to the location where the specimen is kept, the collectionCode refers to the collection where it can be accessed and the catalogNumber should be the number of the specimen or the jar in which it can be accessed.
- **humanObservation**: should be used for sightings of specimens. It is for specimens that were observed and again released.
- **machineObservation**: For data from sensors (tracked or tagged specimens, c-pods,..). Specimens identified by machines (e.g. image recognition and DNA sequencing).
- **livingSpecimen**: an intentionally kept/cultivated living specimen e.g. in an aquarium or culture collection.

occurrenceStatus

This mandatory term needs to be filled out with either "present" or "absent". If your data set only has presences, you will need to add the term occurrenceStatus and fill it with the value "present".

Taxonomy - Introduction to WoRMS and the link with EMODnet Biology

The World Register of Marine Species (**WoRMS**) provides an authoritative and comprehensive list of names of marine organisms, including information on synonymy. The content of WoRMS is controlled by over 250 taxonomic editors. Each of these editors are considered experts for their assigned taxon group and its geographical area.

The register includes only valid names; in other words, names that have been used to describe a taxon in a scientific paper. A species can have many different valid names if the species was described more than once as a species new to science, usually by a different author who was unaware that the species had already been described. These are called synonyms. All synonyms and the accepted names (the name that should be used for the species - usually the name associated to the first description of the species) are included in the register, so that it can serve to standardise the names used in different datasets from different geographical areas at different time periods.

Apart from taxonomic information WoRMS can also store the known distribution area of a taxon, traits information (like the minimum and maximum length of adults), the functional group(s) a taxon belongs to, etc. All this additional information is also controlled by the taxonomic experts.

When dealing with WoRMS, you may see many mentions of APHIA. APHIA is the name of the database in which all WoRMS data are stored. The APHIA database is also used to store some non-marine registers which are not included in the World Register of Marine Species.

By using WoRMS as its taxonomic backbone, EMODnet Biology gains access to all these data stored by WoRMS. For this to be consistent and interoperable, all taxa included in a data set harvested by EMODnet Biology need to reference the Life Sciences Identifier (**LSID**) assigned to the taxon by WoRMS. This information should be stored in the field **scientificNameID**. Names can be easily matched with their correct LSID by using the Taxon Match tool of WoRMS.

If you would like more information on WoRMS, you can watch this online lecture:

.

Taxonomy - WoRMS Taxon Match Tool: preparing an input file

The WoRMS Taxon Match Tool works best if it's presented with clean names. For high efficiency, it is highly recommended to clean your file before presenting it to the taxon match tool.

As explained in a previous section, the main terms to store taxon related data in the occurrence table are scientificName, scientificNameAuthorship, scientificNameID, kingdom, taxonRank and identificationQualifier. For EMODnet Biology, **the field scientificName should only contain the scientific name of the taxon**, other taxonomic information should be split to other columns such as identificationQualifier and scientificNameAuthor. Often the data provided in the name field also contains information related to the sex, lifestage, or size of the specimen. These should be stored in the eMoF table with dedicated terms and if applicable with a reference to a controlled vocabulary. Guidelines on how to correctly occupy the eMoF table will be discussed in the book Data standardisation - eMoF.

Sometimes there is uncertainty about which taxon is actually recorded. This is often expressed by *.cf* followed by the specificEpithet. e.g. *Gadus* cf. *morhua*. In other cases, the person identifing might think the taxon is one of 2 different taxa and will store it as follows: *Gadus morhua / macrocephalus*. The guideline of EMODnet Biology goes as follows: the scientificName with the taxonomic level at which there is certainty should be filled out. So, in both examples the scientificName is filled out with *Gadus*. The identificationQualifier will contain cf. *morhua* and *morhua / macrocephalus* respectively.

In some cases, the name provided is not a taxonomic name, but something else, like a functional group. As mentioned, the field scientificName should only contain scientific names and is a mandatory field. To solve the issue of e.g. functional groups in the name field, a taxon name should be provided instead. Again here, the taxon name that reaches the lowest taxonomic level and still fully covers all taxa implied by the provided name should be chosen. It's better to be overly cautious (and pick a high category taxon name - like animalia or biota) than to introduce errors. The non-taxon names that were originally provided can be stored in the field taxonRemarks.

Examples

The spreadsheet below shows some examples on how to structure your input file before presenting it to the Taxon Match Tool. The field scientificName is the field that is presented to the Taxon Match Tool, and the other columns are needed to add the cleaned data file. The column "Name as provided" is essential to keep in the taxon match file, as this will allow you to link the outcome of the Taxon Match Tool back to your data file.

View in google spreadsheets.

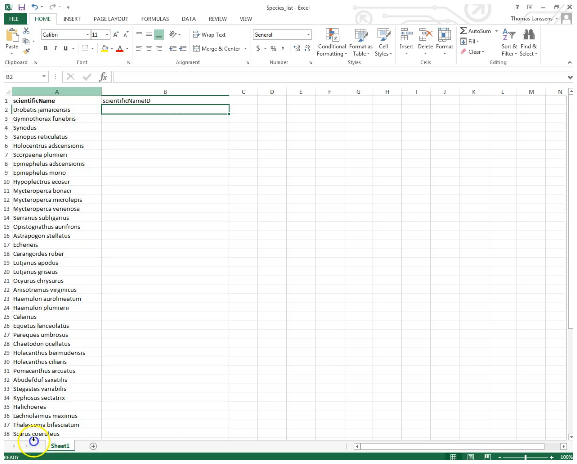
Taxonomy - WoRMS Taxon Match Tool: obtaining the LSIDs

Few taxon records

Single search:

In case your data set only contains 2 or 3 taxa you can quite quickly get their LSIDs from using the search interface.

The following video shows you how you use this search and where to find the LSID:



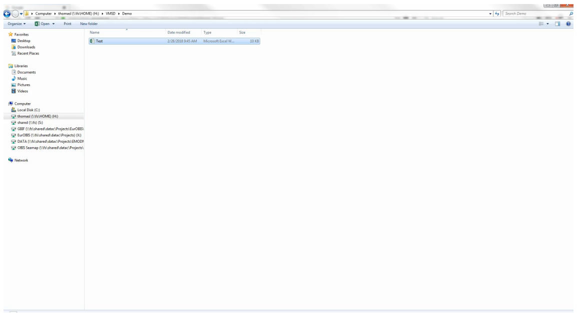
	A	B
1	**scientificName**	scientificNameID
2	Urobatis jamaicensis	
3	Gymnothorax funebris	
4	Synodus	
5	Sanopus reticulatus	
6	Holocentrus adscensionis	
7	Scorpaena plumieri	
8	Epinephelus adscensionis	
9	Epinephelus morio	
10	Hypoplectrus ecosur	
11	Mycteroperca bonaci	
12	Mycteroperca microlepis	
13	Mycteroperca venenosa	
14	Serranus subligarius	
15	Opistognathus aurifrons	
16	Astrapogon stellatus	
17	Echeneis	
18	Carangoides ruber	
19	Lutjanus apodus	
20	Lutjanus griseus	
21	Ocyurus chrysurus	
22	Anisotremus virginicus	
23	Haemulon aurolineatum	
24	Haemulon plumierii	
25	Calamus	
26	Equetus lanceolatus	
27	Pareques umbrosus	
28	Chaetodon ocellatus	
29	Holacanthus bermudensis	
30	Holacanthus ciliaris	
31	Pomacanthus arcuatus	
32	Abudefduf saxatilis	
33	Stegastes variabilis	
34	Kyphosus sectatrix	
35	Halichoeres	
36	Lachnolaimus maximus	
37	Thalassoma bifasciatum	
38	Scarus coeruleus	

Taxon Match Tool:

In case your data set contains many taxa, finding the LSID for each of them will be a tedious task. Therefore, WoRMS has developed a tool which allows easy matching of multiple taxa at the same time; the WoRMS Taxon Match Tool.

A manual for the WoRMS Taxon Match Tool is available here.

The following video shows how you can use this tool to get the LSID's for multiple taxa at the same time.


Computer ▸ thomasl (\\fs\HOME) (H:) ▸ VMSD ▸ Demo
Organize
Open
Print
New folder
Favorites
Desktop
Downloads
Recent Places
Libraries
Documents
Music
Pictures
Videos
Computer
Local Disk (C:)
thomasl (\\fs\HOME) (H:)
shared (\\fs) (S:)
GBIF (\\fs\shared\datac\Projects\EurOBIS\
EurOBIS (\\fs\shared\datac\Projects) (X:)
DATA (\\fs\shared\datac\Projects\EMODN
OBIS Seamap (\\fs\shared\datac\Projects\
Network

Name	Date modified	Type	Size
Test	2/28/2018 9:45 AM	Microsoft Excel W...	10 KB

Features of the Taxon Match Tool:

Fuzzy matches: the Taxon Match Tool takes into account spelling variations of the valid name in WoRMS.

Ambiguous matches: sometimes the same name was used by different authors to describe different species (homonyms). In this case the tool returns the message that the match was ambiguous. Later on we’ll see how to handle these cases.

The **limit to taxa belonging to** helps you reduce ambiguous matches by limiting the available matches to a single taxonomic group.

You can only match 3000 rows at a time. If your data set contains more than 3000 distinct taxa, you can use the LifeWatch web service (see below) or split your distinct taxon list in multiple files and run each of them through the service.

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

Many taxon records

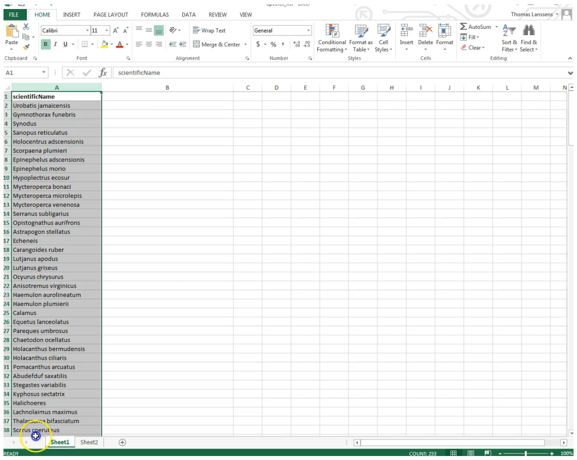
Usually you will have a data set where the same taxonomic name occurs multiple times. A data set can have for example 50,000 occurrence records of only 200 different taxa. In this case it would be much more practical to provide the WoRMS Taxon Match Tool with a distinct list of those 200 taxonomic names that occur in your data set.

To do this you will need to add information from one table to another table. In SQL or R, a join statement can be used, and in Excel the **VLOOKUP** function can be used. This function can be used for adding matched names from the WoRMS taxon matching service to your source data. VLOOKUP accepts the following arguments:

- The cell in the first table which links to the second table
- The range of the second table
- The index of the column in the second table which has the key to link to the first table
- The last argument needs to be FALSE to obtain an exact match

Make sure to use dollar signs to fix the rows and columns in the second argument.

In the following video, we create a worksheet with unique taxon names, submit it to the WoRMS taxon matching service, and add the matched IDs back to our source data using VLOOKUP:



A1 scientificName

	A
1	**scientificName**
2	Urobatis jamaicensis
3	Gymnothorax funebris
4	Synodus
5	Sanopus reticulatus
6	Holocentrus adscensionis
7	Scorpaena plumieri
8	Epinephelus adscensionis
9	Epinephelus morio
10	Hypoplectrus ecosur
11	Mycteroperca bonaci
12	Mycteroperca microlepis
13	Mycteroperca venenosa
14	Serranus subligarius
15	Opistognathus aurifrons
16	Astrapogon stellatus
17	Echeneis
18	Carangoides ruber
19	Lutjanus apodus
20	Lutjanus griseus
21	Ocyurus chrysurus
22	Anisotremus virginicus
23	Haemulon aurolineatum
24	Haemulon plumierii
25	Calamus
26	Equetus lanceolatus
27	Pareques umbrosus
28	Chaetodon ocellatus
29	Holacanthus bermudensis
30	Holacanthus ciliaris
31	Pomacanthus arcuatus
32	Abudefduf saxatilis
33	Stegastes variabilis
34	Kyphosus sectatrix
35	Halichoeres
36	Lachnolaimus maximus
37	Thalassoma bifasciatum
38	Scarus coeruleus

Taxonomy - WoRMS Taxon Match Tool: the incredible return

The Taxon Match Tool returns a file which includes additional information for each taxa that was matched with WoRMS. The fields of relevance that are returned are: Match type, LSID, ScientificName, Kingdom, and isMarine. In the file below, you find the outcome an example taxon match.

View in google spreadsheets.

Match type: This field specifies how precise the match with WoRMS is. “Exact” means that the name in the data file and in WoRMS are exactly the same. If there was no match possible this field remains empty. Values "near_1", "near_2", and "near_3" indicate how many letters difference there are between the name in WoRMS and the name in your data file. Assessing these values seems common sense, but should still be carefully done. Usually, "near_1" indicates a spelling mistake and can reasonably reliably be considered the same taxon (but you should still check!). Matches with type “near_3” should be treated with great caution and should always be checked with the original source or person that identified the taxon. The “phonetic” type indicates that the spelling of the name in the data file and in WoRMS sound the same when spoken, which explains the spelling variations.

ScientificName: This is the scientific name of the taxon as it is recorded in WoRMS. For types with “near” matches, this field should be compared with the original scientific name to see if it's possible that this is indeed the name which was meant.

Authority: If your dataset contains the author of the taxon name, you should check if the Taxon Match Tool returned the same name, especially for the near matches.

Kingdom: All scientific names returned by the taxon match belong to a kingdom. Check if these correspond with the kingdoms you expected; check if there are no animals in a phytoplankton dataset, or fungi in a fish dataset, etc.

isMarine / isBrackish: WoRMS only contains taxa which a taxonomic editor verified as marine or brackish. However, the Taxon Match Tool matches with the APHIA database and can also return matches with non-marine taxa. Checking whether the taxon match returns a marine or brackish taxon is therefore very relevant if your dataset only contains marine taxa. Mismatches can be explained in two ways: 1) the taxon was misidentified in your dataset, 2) the environment is wrongly recorded in WoRMS. Keep in mind that the definition of a marine taxon entails the taxon’s dependency on the marine environment for at least one part of its life cycle; a pigeon which was spotted flying over the ocean is not considered a marine taxon.

LSID: This is what is needed to fill out the DwC field **scientificNameID**. Using the scientificNameID, EMODnet Biology and OBIS can access all information associated to the taxon, including the accepted scientificName, the taxonomy, distribution, and species traits.

Taxonomy - Which information should be included in the dataset for EMODnet?

WoRMS only contains valid names; names that have been described in a scientific publication and that will stay available. However, the status of a taxon name itself can change; a taxon name which is considered accepted today may be considered a junior synonym (and thus not accepted) of another taxon name tomorrow. It's also possible that names which are considered synonyms today may be considered different taxa tomorrow. The only way to keep up with these changes is to provide a link to WoRMS and use the taxonomic information from WoRMS at the moment you assess the data. Therefore, it’s important to use the **LSID of the match** for the field scientificNameID, and not the LSID of the scientificName_accepted.

For the same reason the original name as provided in your dataset should be used to populate the field scientificName. If you are not the party that generated the data, you **should not correct spelling variations without the originator's consent**, as this would reduce the traceability to the original data source and limit the possibility of identifying possible incorrectly added scientificNameID's. Similarly, you **do not add the authority, nor any other DwC classification fields** (kingdom, class, taxonRank) that are returned to you by the Taxon Match Tool. These classification fields should only be filled out if the same information was also provided in the original source dataset.

So, in conclusion: Use the incredible return of the Taxon Match Tool **to verify** the matches with WoRMS, and use **only the LSID** of the matched name to populate the DwC field scientificNameID.

Taxonomy - What if a taxon did not match with WoRMS?

Ambiguous matches

It occurs that different authors used the same name to describe different species (homonyms). As we’ve seen, in these cases the tool returns the message that the match was ambiguous. To decide which of the different LSID's should be used, the name of the authority should be checked. If the authority is not included in the data set, it can be found in the determination guide that was used during the identification of the taxa in the data set. If the authority is not available, the classification should be checked; if for example you are working on a fish data set and one match is a polychaete and the other one is a fish, you know the fish match is the correct one. Another way to check which of the ambiguous matches is correct, the known distribution of the taxa can be consulted in WoRMS or OBIS.
If the data set you are analysing is not your own data set, you should always verify your assumptions with your provider and or ask them to check the determination guide for the authority.

Given these different ambiguous matches, you should **not blindly choose** the occurring accepted name as the correct name. In many cases the accepted name might most likely be correct, but there is no certainty.

Non-marine taxa

If the WoRMS taxon did not find a match, it's possible that the taxon you need is a freshwater or a terrestrial taxon.
These cases can be checked using the Interim Register of Marine and Non-marine Genera (IRMNG) matching services, which are available through https://www.irmng.org/. The IRMNG aims to classify all known genera according to their environment. Matching a taxon with this register will inform you on whether all species of the genus occur exclusively in a marine, brackish, fresh water, terrestrial environment, or in a combination of these 4 environments.

- If the IRMNG lets you know that the taxon you looked for occurs in marine or brackish environment, you should send this taxon name to info@marinespecies.org. This way the WoRMS data management team can contact the relevant taxonomic editor and check whether the taxon can be added to WoRMS.
- If the IRMNG lets you know that the taxon you looked for occurs exclusively in a fresh water or terrestrial environment, you should double check this in your data set. Is it possible for a non-marine species to occur at the mentioned location? You might need to consult your data provider and check if the specimen has not been misidentified. If the data provider claims the taxon actually is marine (i.e. there was a mistake in the IRMNG), you should contact info@marinespecies.org. This way the responsible IRMNG taxonomic editor can assess whether the information in the IRMNG needs to be changed, and whether the taxon should be added to WoRMS.

Note: make sure to keep the environment flag ticked when performing an IRMNG taxon match.

Is it a valid taxon?

Quite often the scientific names of datasets submitted to EMODnet Biology contain spelling variations, or simply invalid names (names which have not been published in literature). When both WoRMS and IRMNG do not provide a match, you should check whether the taxon exists in other registers. The easiest way to do this is by simply searching for the taxon name in google. When this results in 20 hits or less (maybe even referring to your data source), it’s likely that there is something wrong with the taxon name and you should check with your data provider if they actually meant another taxon name. If the google search shows that the taxon exists in other registers such as the ones listed below, you should contact info@marinespecies.org. This way the WoRMS data management team can check whether the taxon can be added to WoRMS.

Examples of trustworthy registers:

- Catalogue of Life – CoL
- Integrated Taxonomic Information System – ITIS
- Pan-European Species-directories Infrastructure – PESI
- International Plant Names Index – IPNI
- Global Names Index - GNI

Developing a tool to check WoRMS taxa versus other registers has been a Belgian Lifewatch project and is available here: https://www.lifewatch.be/data-services, under Select Webservices - **taxon match services**.

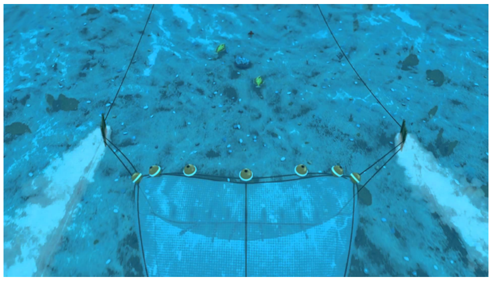
Locations - Exact positions

In case the event has an exact position, the coordinates can be stored in the fields **decimalLatitude** and **decimalLongitude**. The field **coordinateUncertaintyInMeters** can then be used to store the uncertainty associated to the coordinates themselves (e.g. the error of the GPS reading).

These fields (decimalLatitude, decimalLongitude) are mandatory fields. They should contain the geographic latitude and longitude (in decimal degrees), using the spatial reference system EPSG:4326 (WGS84), and the number of decimals should be appropriate for the level of uncertainty related to the precision of the measuring device (this is also reflected in the field coordinateUncertaintyInMeters). Regarding decimalLatitude, positive values are north of the equator, negative values are south of the equator. All values lie between -90 and 90, inclusively. Regarding decimalLongitude, positive values are east of the Greenwich prime meridian, negative values are west of the Greenwich prime meridian. All values lie between -180 and 180, inclusively.

The field **locationID** can be used to store an identifier for monitoring stations.
The field **locality** can be used to store a textual location or a description of where the sample was taken.

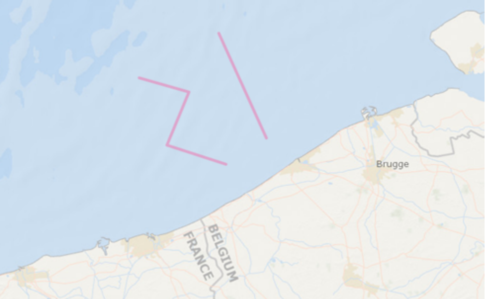
Locations - Start and end positions - Linestrings



In cases where it was not possible to provide an exact position, (e.g. when a sample is collected by trawling along a straight line), it’s recommended to store the start and end positions of the sampling event (in this case the trawl) as a linestring in the field **footprintWKT**. The fields **decimalLatitude** and **decimalLongitude** are still mandatory, but they can be filled out with the centre point of the trawl. The value in **coordinateUncertaintyInMeters** will then represent the distance from the centre to the start or end coordinates.

The linestring should be formatted as follows: "LINESTRING ([start-longitude] [start-latitude], [end-longitude] [end-latitude])". Example: "LINESTRING (2.80151 51.28597, 2.61749 51.53950)".

The OBIS maptool allows you to draw the given linestring on a map after which it will return to you the centre point coordinates and the radius you need to fill out in the field coordinateUncertaintyInMeters.



An excel script allows you to create the linestring and the centre coordinates, to calculate the coordinateUncertaintyInMeters by simply filling out the start and end coordinates.

Locations - General areas and regions - Polygons

If the exact position is unknown, but only the region in which the sample was collected is known, the data might still be valuable and should be included in EMODnet Biology. You could add the name of the region in the field **locality**. Then, you should look up this region in Marine Regions. Marine regions will provide centre coordinates of the region, the coordinateUncertaintyInMeters and the minimum and maximum longitude and latitude. With this information you can build a bounding box polygon for the footprintWKT.

The polygon should be formatted as follows: "POLYGON (([longitude point 1] [latitude point 1], [longitude point 2] [latitude poin 2], ..., [longitude point x] [latitude point x], [longitude point 1] [latitude point 1]))". Example: "POLYGON ((10.65674 42.77928, 10.50018 42.77121, 10.43152 42.62183, 10.75836 42.38087, 11.05225 42.48628, 10.91492 42.70262, 10.65674 42.77928))".

Alternatively, if you know approximately where the sample was collected you may opt to draw your own polygon using the OBIS maptool as illustrated in the demo video below.

If you are skilled in R, you can use these R tools to calculate the centroid and radius for WKT linestrings and polygons.
No tool exists in Python to date.

Data structure

This topic focuses on structuring data in Event Core and Occurrence Core formats. A transformation workflow on how to fill in the three different tables (event, occurrence, emof) is provided.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Data structure
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:28 PM

Table of contents

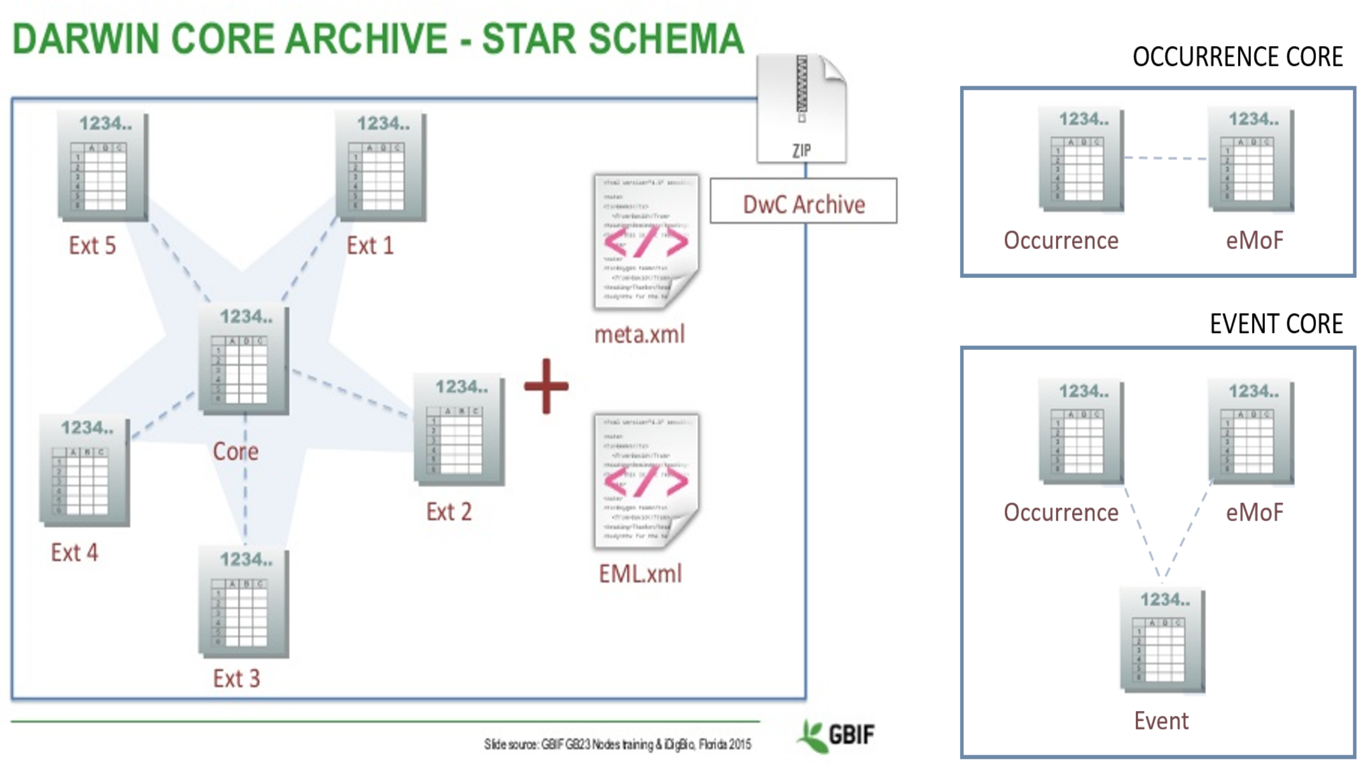
Introduction

When to use Event Core and when to use Occurrence Core

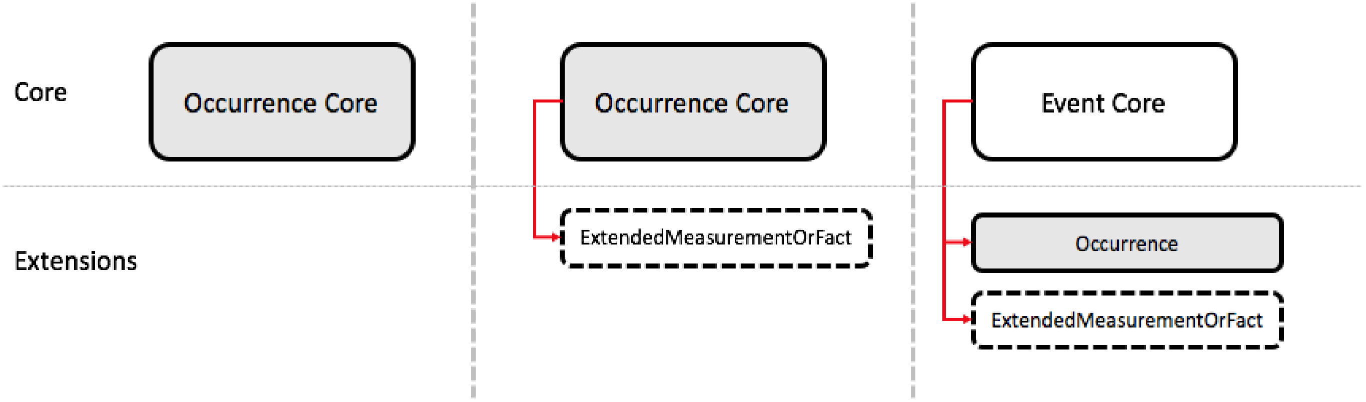
Transformation workflow
- Creating the Occurrence table
- Creating the Event table
- Normalising the Event table and adding a hierarchy
- Creating the eMoF table

Introduction

Darwin Core Archive (DwC-A) is the standard for publishing biodiversity data using Darwin Core terms. As we've seen, this model is used in EMODnet Biology and OBIS (and GBIF). The conceptual data model of the Darwin Core Archive is a “star schema” with a core table in the center of the star and extension tables radiating out of the center, linked by database keys such as ID columns. In practice, EMODnet Biology and OBIS use a subset of 1 to 3 tables to represent the data. In most cases, we will use the three tables.



Slide source: GBIF GB23 Nodes training & iDigBio, Florida 2015



Reminder: What are the tables used for?

Event table:

- to store sample and/or observation information (time, location, depth, event hierarchy)

Occurrence table:

- to store occurrence details (taxonomy, identification, organismID...)

Extended Measurements or Facts (eMoF) table:

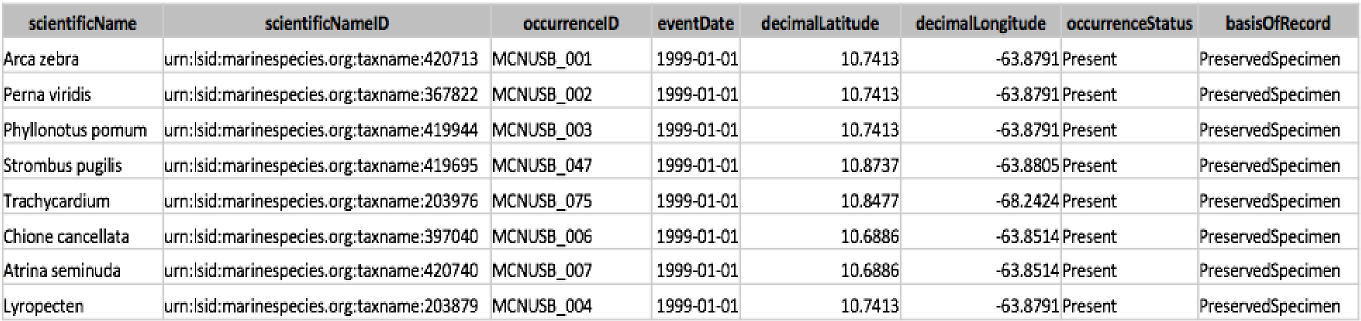
- organism quantifications (e.g. counts, abundance, biomass, % live cover, etc.)
- species biometrics (e.g. body length, weight, etc.)
- facts documenting a specimen (e.g. living/dead, behaviour, invasiveness, etc.)
- abiotic measurements (e.g. temperature, salinity, oxygen, sediment grain size, habitat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

When to use Event Core and when to use Occurrence Core

Depending on the nature of your data, you can organize your dataset in three different ways: *Occurrence Core without ExtendedMeasurementsOrFact (eMoF) extension*, *Occurrence Core with eMoF extension* or *Event Core with Occurrence Core extension and ExtendedMeasurementOrFact extension.*

Occurrence Core without eMoF:

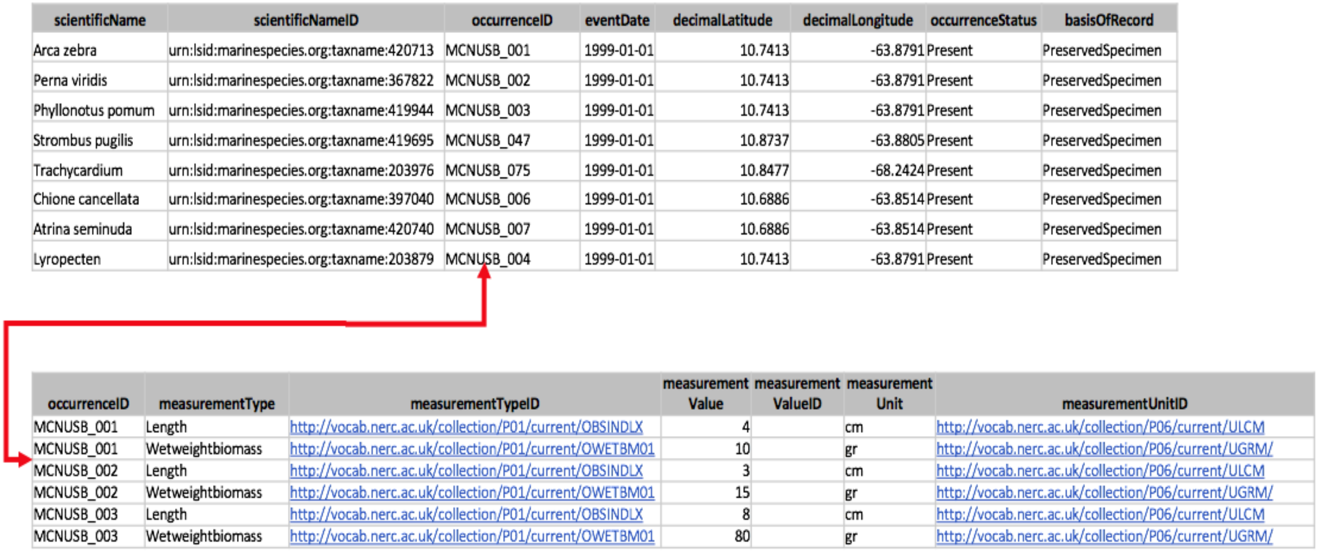
- There is no information on how the data was sampled or samples were processed.
- There are no abiotic measurements taken or provided
- Biological measurements are made on individual specimens (each specimen is a single occurrence record)
- This is often the case for museum collections, citations of occurrences from literature, individual sightings.
- The occurrenceID is the unique identifier in this table.



scientificName	scientificNameID	occurrenceID	eventDate	decimalLatitude	decimalLongitude	occurrenceStatus	basisOfRecord
Arca zebra	urn:lsid:marinespecies.org:taxname:420713	MCNUSB_001	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen
Perna viridis	urn:lsid:marinespecies.org:taxname:367822	MCNUSB_002	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen
Phyllonotus pomum	urn:lsid:marinespecies.org:taxname:419944	MCNUSB_003	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen
Strombus pugilis	urn:lsid:marinespecies.org:taxname:419695	MCNUSB_047	1999-01-01	10.8737	-63.8805	Present	PreservedSpecimen
Trachycardium	urn:lsid:marinespecies.org:taxname:203976	MCNUSB_075	1999-01-01	10.8477	-68.2424	Present	PreservedSpecimen
Chione cancellata	urn:lsid:marinespecies.org:taxname:397040	MCNUSB_006	1999-01-01	10.6886	-63.8514	Present	PreservedSpecimen
Atrina seminuda	urn:lsid:marinespecies.org:taxname:420740	MCNUSB_007	1999-01-01	10.6886	-63.8514	Present	PreservedSpecimen
Lyropecten	urn:lsid:marinespecies.org:taxname:203879	MCNUSB_004	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen

Occurrence Core with eMoF:

- Occurrences and measurements made on individual specimens, e.g. body size, counts, wet weight, life stage, etc.
- The data in the two tables are connected by their occurrenceID.

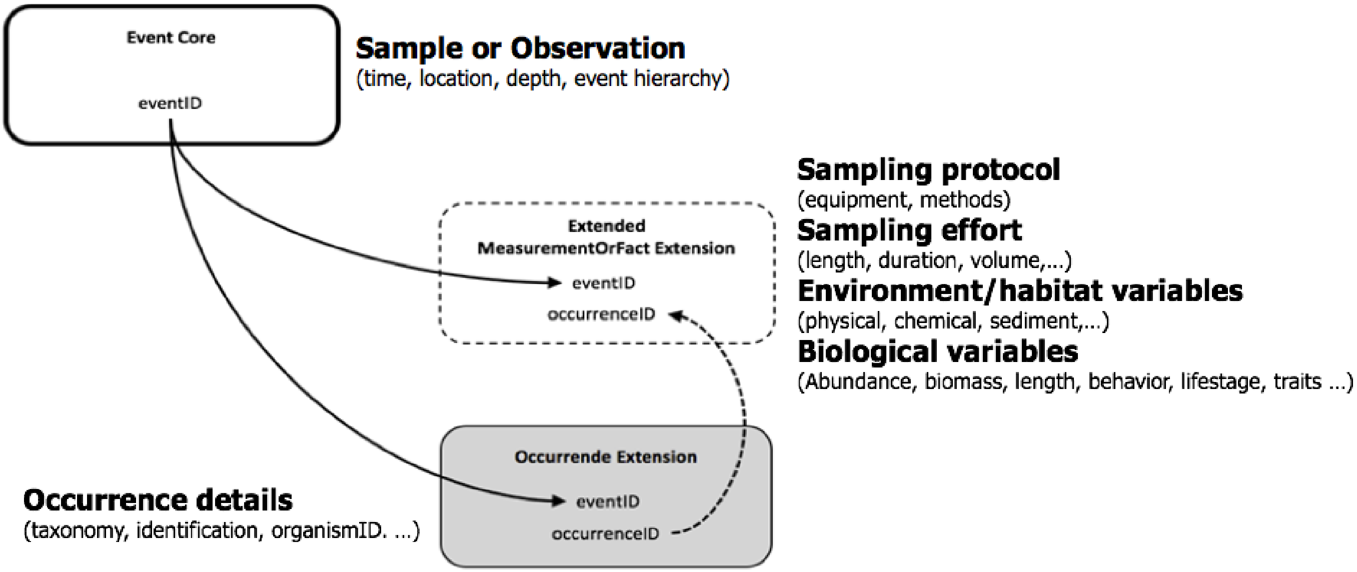


scientificName	scientificNameID	occurrenceID	eventDate	decimalLatitude	decimalLongitude	occurrenceStatus	basisOfRecord
Arca zebra	urn:lsid:marinespecies.org:taxname:420713	MCNUSB_001	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen
Perna viridis	urn:lsid:marinespecies.org:taxname:367822	MCNUSB_002	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen
Phyllonotus pomum	urn:lsid:marinespecies.org:taxname:419944	MCNUSB_003	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen
Strombus pugilis	urn:lsid:marinespecies.org:taxname:419695	MCNUSB_047	1999-01-01	10.8737	-63.8805	Present	PreservedSpecimen
Trachycardium	urn:lsid:marinespecies.org:taxname:203976	MCNUSB_075	1999-01-01	10.8477	-68.2424	Present	PreservedSpecimen
Chione cancellata	urn:lsid:marinespecies.org:taxname:397040	MCNUSB_006	1999-01-01	10.6886	-63.8514	Present	PreservedSpecimen
Atrina seminuda	urn:lsid:marinespecies.org:taxname:420740	MCNUSB_007	1999-01-01	10.6886	-63.8514	Present	PreservedSpecimen
Lyropecten	urn:lsid:marinespecies.org:taxname:203879	MCNUSB_004	1999-01-01	10.7413	-63.8791	Present	PreservedSpecimen

occurrenceID	measurementType	measurementTypeID	measurement Value	measurement ValueID	measurement Unit	measurementUnitID
MCNUSB_001	Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX	4		cm	http://vocab.nerc.ac.uk/collection/P06/current/ULCM
MCNUSB_001	Wetweightbiomass	http://vocab.nerc.ac.uk/collection/P01/current/OWETBM01	10		gr	http://vocab.nerc.ac.uk/collection/P06/current/UGRM/
MCNUSB_002	Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX	3		cm	http://vocab.nerc.ac.uk/collection/P06/current/ULCM
MCNUSB_002	Wetweightbiomass	http://vocab.nerc.ac.uk/collection/P01/current/OWETBM01	15		gr	http://vocab.nerc.ac.uk/collection/P06/current/UGRM/
MCNUSB_003	Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX	8		cm	http://vocab.nerc.ac.uk/collection/P06/current/ULCM
MCNUSB_003	Wetweightbiomass	http://vocab.nerc.ac.uk/collection/P01/current/OWETBM01	80		gr	http://vocab.nerc.ac.uk/collection/P06/current/UGRM/

Event Core with Occurrence Core extension and extendedMeasurementOrFact extension:

- When the data set contains abiotic measurements, or other biological measurements which are related to an entire sample (not a single specimen)
- When specific details are known about how a biological sample was taken and processed.
- Event Core should be used in combination with the Occurrence Extension and the ExtendedMeasurementOrFact Extension.
- The data in the three tables are connected by their occurrenceID and/or eventID.


Event Core
eventID
Sample or Observation
(time, location, depth, event hierarchy)
Extended
MeasurementOrFact Extension
eventID
occurrenceID
Sampling protocol
(equipment, methods)
Sampling effort
(length, duration, volume,...)
Environment/habitat variables
(physical, chemical, sediment,...)
Biological variables
(Abundance, biomass, length, behavior, lifestage, traits ...)
Occurrende Extension
eventID
occurrenceID
Occurrence details
(taxonomy, identification, organismID. ...)

Transformation workflow

This chapter covers the workflow of splitting a flat table into our three DwC tables; event, occurrence, and eMoF. This is necessary to integrate the data in the EurOBIS relational database.

The summarised workflow is as follows:

1. **Occurrence table**: Select the relevant columns/fields of the Occurrence table and place them in a different sheet/table that is named “Occurrence”. At this moment it is useful to include the fields that contain biological measurements related to the occurrences (e.g. biomass).
2. **Event table**: Select the relevant columns/fields of the Event table and place them in a different sheet/table that is named “Event”. At this moment it is useful to include the fields that contain abiotic measurements and other information related to the events (e.g. sampling protocol).
3. **Normalise Events** table and add a hierarchy if relevant.
4. **eMoF table**: Add the necessary columns to the eMoF table and put all the according values into one single column.

In the next sub-chapters you will go through each step.

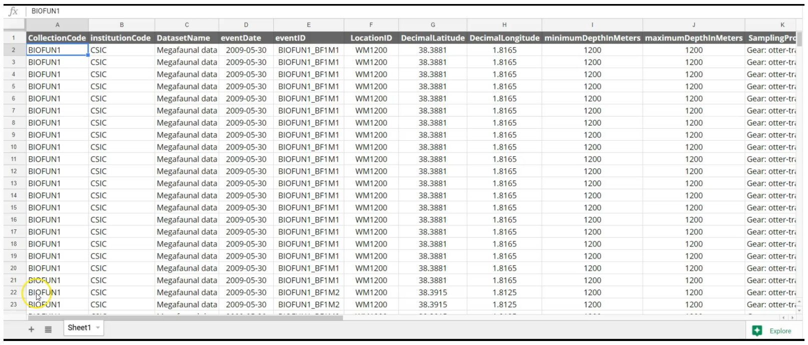
What are the relevant columns?

In the book on Field nomenclature we've mapped the column names of our data set to DwC terms. This allows us to have a better overview on which columns belong to which table.

If you need a reminder on the most relevant DwC fields per table you can find them here.

Creating the Occurrence table

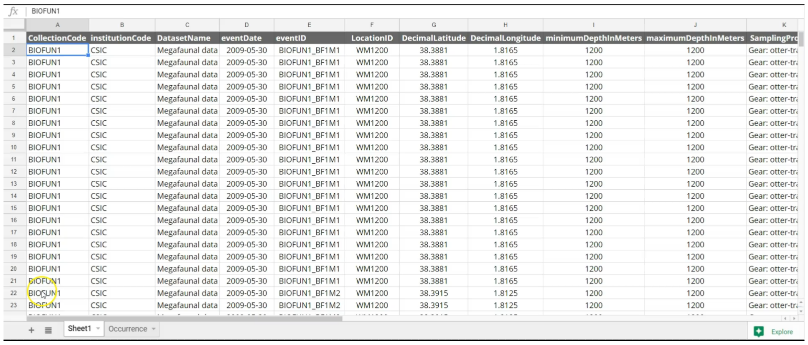
In the video below, you can see how to select the relevant columns of the Occurrence table. At this stage it is helpful to include the measurements or fields that contain biological measurements related to the Occurrences (e.g. biomass).



	A	B	C	D	E	F	G	H	I	J	K
1	CollectionCode	institutionCode	DatasetName	eventDate	eventID	LocationID	DecimalLatitude	DecimalLongitude	minimumDepthInMeters	maximumDepthInMeters	SamplingPro
2	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
3	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
4	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
5	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
6	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
7	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
8	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
9	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
10	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
11	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
12	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
13	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
14	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
15	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
16	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
17	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
18	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
19	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
20	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
21	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
22	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	1200	Gear: otter-tra
23	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	1200	Gear: otter-tra

Creating the Event table

In the video below, you can see how to select the relevant columns of the Event table. At this stage it is helpful to include abiotic measurements (if available) and other information related to the sampling event (e.g. SamplingProtocol).

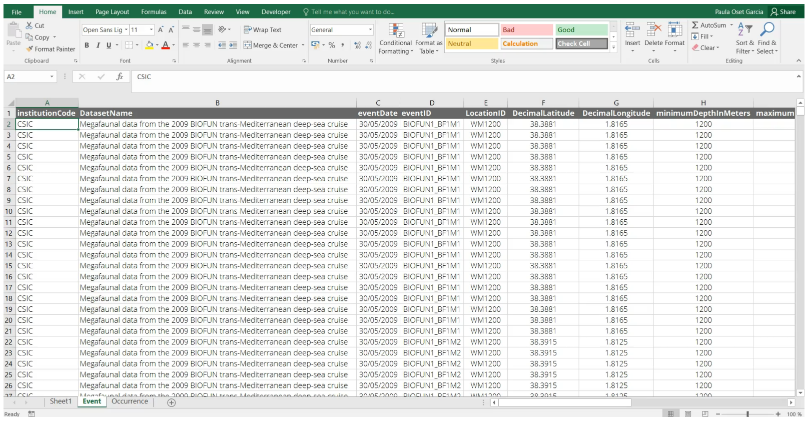


	A	B	C	D	E	F	G	H	I	J	K
1	CollectionCode	institutionCode	DatasetName	eventDate	eventID	LocationID	DecimalLatitude	DecimalLongitude	minimumDepthInMeters	maximumDepthInMeters	SamplingPro
2	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
3	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
4	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
5	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
6	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
7	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
8	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
9	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
10	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
11	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
12	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
13	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
14	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
15	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
16	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
17	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
18	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
19	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
20	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
21	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-tra
22	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	1200	Gear: otter-tra
23	BIOFUN1	CSIC	Megafaunal data	2009-05-30	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	1200	Gear: otter-tra

Normalising the Event table and adding a hierarchy

To normalise the Event table, we will restructure it.

As mentioned, the necessary fields for the Event table are: CollectionCode, EventDate, EventID, LocationID, DecimalLatitude, DecimalLongitude, minimumDepthInMeters, maximumDepthInMeters, SamplingProtocol.
By leaving ScientificName out of the selection, we are left with a table that contains many duplicates. We need to remove the duplicates to create a table with only **unique Events**. We can do this by using the "Remove duplicates" function in Excel:



	A	B	C	D	E	F	G	H	
1	institutionCode	DatasetName	eventDate	eventID	LocationID	DecimalLatitude	DecimalLongitude	minimumDepthInMeters	maximum
2	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
3	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
4	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
5	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
6	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
7	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
8	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
9	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
10	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
11	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
12	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
13	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
14	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
15	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
16	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
17	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
18	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
19	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
20	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
21	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	WM1200	38.3881	1.8165	1200	
22	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	
23	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	
24	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	
25	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	
26	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M2	WM1200	38.3915	1.8125	1200	

Creating the eMoF table

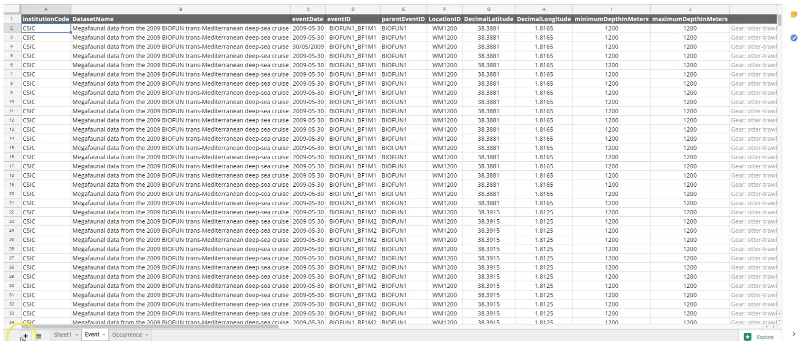
Two basic principles

There are two particularities about the eMoF table that you need to know about before you create it:

- We need to distinguish between data related to the occurrences (e.g. biomass, abundance and count) and data realated to sampling events (e.g. sampling instrument). We will treat them slightly differently.
  As you know, the eMoF table contains both the fields eventID and occurrenceID. For the biotic measurements related to the Occurrences (CountInSample, Abundance (N/km2), and Wet Weight Biomass (kg/km2) in this case), we will need to copy both the eventID and the occurrenceID. For the abiotic measurements or other data related to the sampling event, we will only copy the eventID column.
- We will only use **one column** to store all the measurement values. In the video example, the values now stored in 4 columns will be placed into one single column (measurementValue). The field name as we know it now, will be stored as a value of the column measurementType. This means we have to copy-paste these columns one by one.

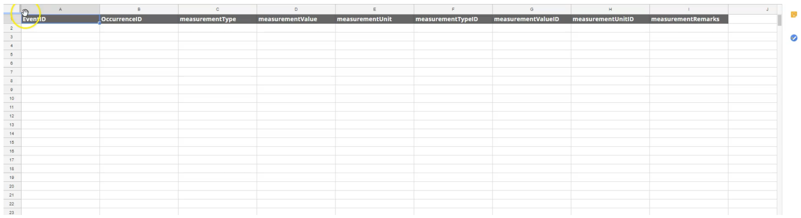
Workflow

1. Create a new sheet with all the necessary eMoF columns: EventID, OccurrenceID, measurementType, measurementValue, measurementUnit, measurementTypeID, measurementValueID, measurementUnitID and measurementRemarks.

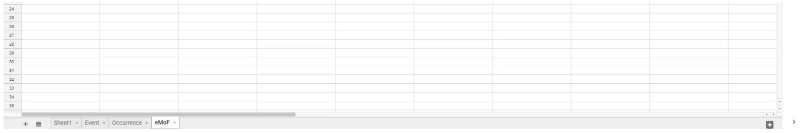


	institutionCode	DatasetName	eventDate	eventID	parentEventID	LocationID	DecimalLatitude	DecimalLongitude	minimumDepthInMeters	maximumDepthInMeters	
2	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
3	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
4	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	30/05/2009	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
5	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
6	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
7	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
8	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
9	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
10	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
11	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
12	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
13	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
14	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
15	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
16	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
17	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
18	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
19	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
20	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
21	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M1	BIOFUN1	WM1200	38.3881	1.8165	1200	1200	Gear: otter-trawl
22	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
23	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
24	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
25	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
26	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
27	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
28	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
29	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
30	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
31	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
32	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
33	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl
34	CSIC	Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise	2009-05-30	BIOFUN1_BF1M2	BIOFUN1	WM1200	38.3915	1.8125	1200	1200	Gear: otter-trawl

2. Add biotic measurements, the example begins with the CountInSample column. Include the the necessary values of EventID, OccurrenceID and CountInSample columns, and fill out all the rows of the column [eMoF.measurementType] with "Count in sample".

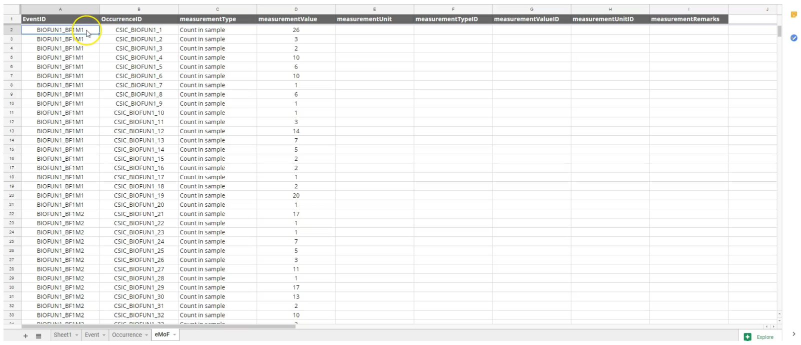


EventID	OccurrenceID	measurementType	measurementValue	measurementUnit	measurementTypeID	measurementValueID	measurementUnitID	measurementRemarks



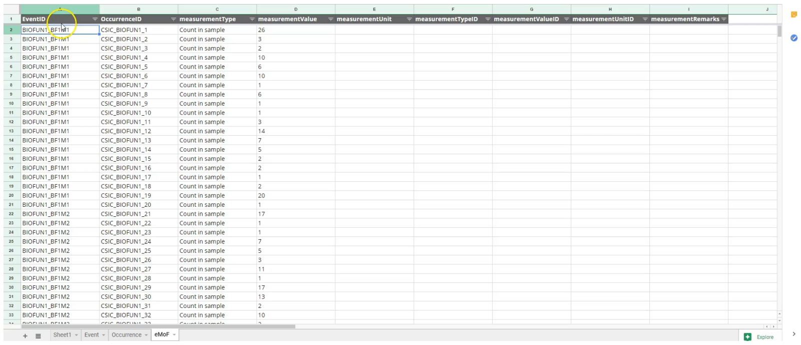
3. Add other biotic measurements. E.g.: Abundance (N/km2). Include the the necessary values of EventID, OccurrenceID and CountInSample columns, and fill out all the rows of the column [eMoF.measurementType] with “Abundance” and fill in the corresponding rows of the column [eMoF.measurementUnit] with “N/km2”.

Repeat for Wet Weight Biomass (kg/km2), making sure to fill in measurementType and measurementUnit with the appropriate values: “Wet weight biomass” and “kg/km2”.



EventID	OccurrenceID	measurementType	measurementValue	measurementUnit	measurementTypeID	measurementValueID	measurementUnitID	measurementRemarks
BIOFUN1_BF1M1	CSIC_BIOFUN1_1	Count in sample	26					
BIOFUN1_BF1M1	CSIC_BIOFUN1_2	Count in sample	3					
BIOFUN1_BF1M1	CSIC_BIOFUN1_3	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_4	Count in sample	10					
BIOFUN1_BF1M1	CSIC_BIOFUN1_5	Count in sample	6					
BIOFUN1_BF1M1	CSIC_BIOFUN1_6	Count in sample	10					
BIOFUN1_BF1M1	CSIC_BIOFUN1_7	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_8	Count in sample	6					
BIOFUN1_BF1M1	CSIC_BIOFUN1_9	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_10	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_11	Count in sample	3					
BIOFUN1_BF1M1	CSIC_BIOFUN1_12	Count in sample	14					
BIOFUN1_BF1M1	CSIC_BIOFUN1_13	Count in sample	7					
BIOFUN1_BF1M1	CSIC_BIOFUN1_14	Count in sample	5					
BIOFUN1_BF1M1	CSIC_BIOFUN1_15	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_16	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_17	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_18	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_19	Count in sample	20					
BIOFUN1_BF1M1	CSIC_BIOFUN1_20	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_21	Count in sample	17					
BIOFUN1_BF1M2	CSIC_BIOFUN1_22	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_23	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_24	Count in sample	7					
BIOFUN1_BF1M2	CSIC_BIOFUN1_25	Count in sample	5					
BIOFUN1_BF1M2	CSIC_BIOFUN1_26	Count in sample	3					
BIOFUN1_BF1M2	CSIC_BIOFUN1_27	Count in sample	11					
BIOFUN1_BF1M2	CSIC_BIOFUN1_28	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_29	Count in sample	17					
BIOFUN1_BF1M2	CSIC_BIOFUN1_30	Count in sample	13					
BIOFUN1_BF1M2	CSIC_BIOFUN1_31	Count in sample	2					
BIOFUN1_BF1M2	CSIC_BIOFUN1_32	Count in sample	10					

4. Add the samplingProtocol field. You have stored this in the Event table because it's related to the sampling event and not to the occurrences. For this type of measurements, that apply to the whole sample, we only need to include the eventID in the eMoF table. For example, if we collect data on the sediment temperature of a benthos sample, this temperature value applies to the whole sample and, therefore, relates to all the Occurrences that belong to that sample, which share the same eventID. By linking sampling and abiotic measurements directly to the eventID and by having our event table already normalised, we avoid duplication of data. Copy the values of the events and samplingProtocol columns, and place them into [eMoF.eventID] and [eMoF.measurementValue] respectively, under the previous records. Fill in all the blank rows of the column [eMoF.measurementType] with "Sampling protocol".



EventID	OccurrenceID	measurementType	measurementValue	measurementUnit	measurementTypeID	measurementValueID	measurementUnitID	measurementRemarks
BIOFUN1_BF1M1	CSIC_BIOFUN1_1	Count in sample	26					
BIOFUN1_BF1M1	CSIC_BIOFUN1_2	Count in sample	3					
BIOFUN1_BF1M1	CSIC_BIOFUN1_3	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_4	Count in sample	10					
BIOFUN1_BF1M1	CSIC_BIOFUN1_5	Count in sample	6					
BIOFUN1_BF1M1	CSIC_BIOFUN1_6	Count in sample	10					
BIOFUN1_BF1M1	CSIC_BIOFUN1_7	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_8	Count in sample	6					
BIOFUN1_BF1M1	CSIC_BIOFUN1_9	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_10	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_11	Count in sample	3					
BIOFUN1_BF1M1	CSIC_BIOFUN1_12	Count in sample	14					
BIOFUN1_BF1M1	CSIC_BIOFUN1_13	Count in sample	7					
BIOFUN1_BF1M1	CSIC_BIOFUN1_14	Count in sample	5					
BIOFUN1_BF1M1	CSIC_BIOFUN1_15	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_16	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_17	Count in sample	1					
BIOFUN1_BF1M1	CSIC_BIOFUN1_18	Count in sample	2					
BIOFUN1_BF1M1	CSIC_BIOFUN1_19	Count in sample	20					
BIOFUN1_BF1M1	CSIC_BIOFUN1_20	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_21	Count in sample	17					
BIOFUN1_BF1M2	CSIC_BIOFUN1_22	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_23	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_24	Count in sample	7					
BIOFUN1_BF1M2	CSIC_BIOFUN1_25	Count in sample	5					
BIOFUN1_BF1M2	CSIC_BIOFUN1_26	Count in sample	3					
BIOFUN1_BF1M2	CSIC_BIOFUN1_27	Count in sample	11					
BIOFUN1_BF1M2	CSIC_BIOFUN1_28	Count in sample	1					
BIOFUN1_BF1M2	CSIC_BIOFUN1_29	Count in sample	17					
BIOFUN1_BF1M2	CSIC_BIOFUN1_30	Count in sample	13					
BIOFUN1_BF1M2	CSIC_BIOFUN1_31	Count in sample	2					
BIOFUN1_BF1M2	CSIC_BIOFUN1_32	Count in sample	10					

At this stage, the data are completely structured. The event table is normalised and linked to the occurrence and eMoF table by an eventID. The occurrence table is linked to the measurements in the eMoF table by an occurrenceID.
In a next chapter, we will go through the workflow of standardising the content of the eMoF table with controlled vocabularies.

Data standardisation - eMoF

This topic provides detailed information on how to increase the interoperability of the extended Measurement or Fact extension table by using controlled vocabularies from BODC

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Data standardisation - eMoF
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:29 PM

Table of contents

eMoF table and controlled vocabularies

Introduction to BODC controlled vocabularies

Choosing your BODC vocab

Dealing with measurementTypes: Biotic measurements or facts

Dealing with measurementTypes: Abiotic measurements or facts

Dealing with measurementTypes: Sampling facts

Dealing with measurementValues

Dealing with measurementUnits

eMoF table and controlled vocabularies

Previously you have seen how to structure the data into Event Core format and we have started populating the eMoF table.

For the moment, we have only used the **measurementType**, **measurementValue** and **measurementUnit** fields. These are completely unconstrained, human readable fields and can be populated with free text annotation. The main advantage of free text fields is that it allows to capture complex and yet unclassified information. However, free text introduces heterogeneity (e.g. spelling or wording differences) that become a major challenge for effective data integration and analysis.

The eMoF table provides three additional fields to standardise the measurement types, values and units: **measurementTypeID**, **measurementValueID** and **measurementUnitID**. These are the columns that will be populated using controlled vocabularies from the Natural Environment Research Council (NERC) Vocabulary Server, developed, maintained and governed by the British Oceanographic Data Centre (BODC).

Example of parameter standardisation using controlled vocabularies:

MeasurementType (free text)	**MeasurementTypeID** (controlled vocabulary)
Body length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
Length	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
Length (mm)	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
length_in_mm	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX
Length of specimen	http://vocab.nerc.ac.uk/collection/P01/current/OBSINDLX

Note: although measurementType, measurementValue and measurementUnit are free text fields, it is recommended to fill them in with the "preferred label" given by the BODC parameter.

Introduction to BODC controlled vocabularies

Controlled vocabularies are collections of concepts that can be used to populate a given field in a data or metadata model. The BODC vocabularies comprise several collections (also called lists) of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic (and wider) community. The vocabulary services are technically managed and hosted by the British Oceanographic Data Centre (BODC) by means of the NERC Vocabulary Server

Using standardised sets of terms solves the problem of ambiguities associated with data markup and also enables records to be interpreted by computers. This opens up datasets to a whole world of possibilities for computer aided manipulation, distribution and long term reuse. See more info here.

In the eMoF table, you use the measurementTypeID, measurementValueID and measurementUnitID fields to store a machine readable label (i.e. a code, URI or URL).

But where can you find these concepts?

There are different interfaces to search for these concepts and their URIs. For example:

- The BODC vocabulary search
- The SeaDataNet Facet search (for the P01 collection)

Finding the appropriate vocabulary (and, therefore, which URI to include in the eMoF table) can be challenging without some insight and understanding of the semantic model behind. This falls outside the scope of this course. However, what is interesting to know is that all the concept URIs that you need to provide follow this structure:

“http://vocab. nerc.ac.uk/collection/{listID}/{version}/{termID}/”

Examples

- http://vocab.nerc.ac.uk/collection/P01/current/ADBIOL01/
    - {listID} = P01 -> this means we are looking into the collection known as P01 or "Parameter usage vocabulary"
    - {version} = current
    - {termID} = ADBIOL01 -> an ID for this concept (Abundance of biological entity specified elsewhere per unit length sampled of the water body)
- http://vocab.nerc.ac.uk/collection/P06/current/KGXX/
    - {listid} = P06 -> this means we are looking into the collection known as P06 or "BODC-approved data storage units"
    - {version} = current
    - {termid} = KGXX -> an ID for this concept (Kilograms)

Choosing your BODC vocab

BODC contains a large amount of terms and sometimes you can find into two vocabs that might look identical. In these cases you will need to choose the one that applies the best to your data (if any). If the term you are looking for does not exist yet in BODC, it can be requested (contact bio@emodnet.eu for more information).

There are a few tips that will help us deciding which BODC vocab to use:

- **Vocab term suitability**: Use a BODC vocab only if it has the exact same meaning as your term, e.g.:
  - If your measurement is "*Concentration of phosphate in the water body*", you should rather add the BODC parameter http://vocab.nerc.ac.uk/collection/P35/current/EPC00007/ and not just http://vocab.nerc.ac.uk/collection/S27/current/CS026904/

- **Collection suitability**: Use a BODC vocab from a collection that is adequate for your data, e.g.:
  - If your fact is "Sex/gender of an organism", searching for "sex" in the BODC Vocabulary Search would result in 17 different collections. Let's say after a quick search you are hesitating between http://vocab.nerc.ac.uk/collection/P01/current/ENTSEX01/ from the "P01 BODC Parameter Usage Vocabulary collection" and http://vocab.nerc.ac.uk/collection/MVB/current/MVB000023/ from the "MVB Movebank Attribute Dictionary collection".
    To assess which collection to use, look carefully at the title and definition of the collections to see which one of them relates the most to the type of data you are handling:

Title	**Definition**
P01 BODC Parameter Usage Vocabulary	*Terms built using the BODC parameter semantic model designed to describe individual measured phenomena. May be used to mark up sets of data such as a NetCDF array or spreadsheet column.*
MVB Movebank Attribute Dictionary	*Terms used to describe on-animal sensor data stored in the Movebank database (movebank.org), including individual measurements (events), reference data (animals, tags, deployments) and studies.*

Pro Tip: *Access the whole collection by using a URL of the type https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/{listid}/ -> https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/P06/*

OR click on search without specifying a collection nor a "search text" to access the metadata of all the collections. Download the results for a better check

- **Vocab term integrity**: Make sure that your term is not deprecated and that you are using the most recent version of the vocab term. Using 'Simple search within a vocabulary' in https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/ is the best and fastest way to browse and find non-deprecated current terms.

To make things easier, the most often used BODC collections within OBIS-ENV format are identified below:

measurementTypes

- Q01- OBIS sampling instruments and methods attributes
- P01- BODC Parameter Usage Vocabulary

measurementValues

- 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

measurementUnits

- P06- Approved data storage units

Dealing with measurementTypes: Biotic measurements or facts

Searching your concept

We will provide you with a filtered list of P01 parameters that should be used for the biotic measurements (or facts) that your dataset contains (e.g. biomass, abundance, sex/gender, life or development stage, etc.).

To access the full list search for "biological%entity%specified%elsewhere" within the P01 BODC Parameter Usage Vocabulary collection

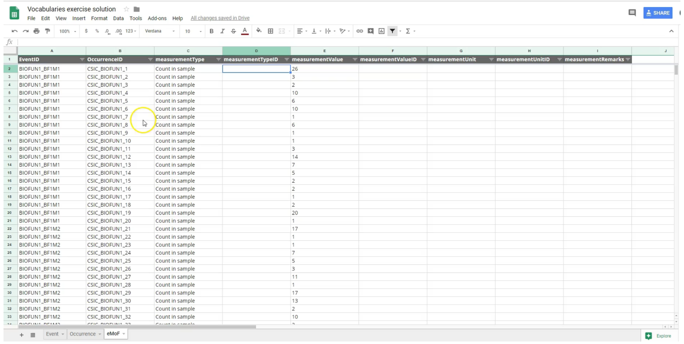
As you can see, this is a search result using the BODC Vocabulary Search. We have searched for all the concepts that contain the free text: "biological entity specified elsewhere". In our case, the biological entity in question is specified in the Occurrence table, linked to the measurement via the OccurrenceID.

The same search could have been performed using the SDN Facet search tool. The P01 collection that you will use to populate many of the measurementTypeIDs is quite interesting. Click in the following link if you want to know more about P01 BODC Parameter Usage Vocabulary context and semantic model.

Now, look at the list and select the most appropriate parameter for our measurements "Count in sample", "abundance (N/km2)" and "Biomass (kg/km2)".

Populating measurementTypeID

Once you have found the appropriate BODC parameters, you can use them in your eMoF table as is shown in the following video:


Vocabularies exercise solution
File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive
SHARE
100% $ % .0 .00 123 Verdana 10
Event
Occurrence
eMoF
Explore

EventID	OccurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID	measurementRemarks
BIOFUN1_BF1M1	CSIC_BIOFUN1_1	Count in sample		26				
BIOFUN1_BF1M1	CSIC_BIOFUN1_2	Count in sample		3				
BIOFUN1_BF1M1	CSIC_BIOFUN1_3	Count in sample		2				
BIOFUN1_BF1M1	CSIC_BIOFUN1_4	Count in sample		10				
BIOFUN1_BF1M1	CSIC_BIOFUN1_5	Count in sample		6				
BIOFUN1_BF1M1	CSIC_BIOFUN1_6	Count in sample		10				
BIOFUN1_BF1M1	CSIC_BIOFUN1_7	Count in sample		1				
BIOFUN1_BF1M1	CSIC_BIOFUN1_8	Count in sample		6				
BIOFUN1_BF1M1	CSIC_BIOFUN1_9	Count in sample		1				
BIOFUN1_BF1M1	CSIC_BIOFUN1_10	Count in sample		1				
BIOFUN1_BF1M1	CSIC_BIOFUN1_11	Count in sample		3				
BIOFUN1_BF1M1	CSIC_BIOFUN1_12	Count in sample		14				
BIOFUN1_BF1M1	CSIC_BIOFUN1_13	Count in sample		7				
BIOFUN1_BF1M1	CSIC_BIOFUN1_14	Count in sample		5				
BIOFUN1_BF1M1	CSIC_BIOFUN1_15	Count in sample		2				
BIOFUN1_BF1M1	CSIC_BIOFUN1_16	Count in sample		2				
BIOFUN1_BF1M1	CSIC_BIOFUN1_17	Count in sample		1				
BIOFUN1_BF1M1	CSIC_BIOFUN1_18	Count in sample		2				
BIOFUN1_BF1M1	CSIC_BIOFUN1_19	Count in sample		20				
BIOFUN1_BF1M1	CSIC_BIOFUN1_20	Count in sample		1				
BIOFUN1_BF1M2	CSIC_BIOFUN1_21	Count in sample		17				
BIOFUN1_BF1M2	CSIC_BIOFUN1_22	Count in sample		1				
BIOFUN1_BF1M2	CSIC_BIOFUN1_23	Count in sample		1				
BIOFUN1_BF1M2	CSIC_BIOFUN1_24	Count in sample		7				
BIOFUN1_BF1M2	CSIC_BIOFUN1_25	Count in sample		5				
BIOFUN1_BF1M2	CSIC_BIOFUN1_26	Count in sample		3				
BIOFUN1_BF1M2	CSIC_BIOFUN1_27	Count in sample		11				
BIOFUN1_BF1M2	CSIC_BIOFUN1_28	Count in sample		1				
BIOFUN1_BF1M2	CSIC_BIOFUN1_29	Count in sample		17				
BIOFUN1_BF1M2	CSIC_BIOFUN1_30	Count in sample		13				
BIOFUN1_BF1M2	CSIC_BIOFUN1_31	Count in sample		2				
BIOFUN1_BF1M2	CSIC_BIOFUN1_32	Count in sample		10				

The procedure would consist in copying the URI for each concept and pasting it into the measurementTypeID field for the records that correspond to that measurementType.

Dealing with measurementTypes: Abiotic measurements or facts

Let's suppose that your data also contains abiotic measurements that were collected at the moment of sampling (e.g. salinity of the water body, temperature of the water body or sediment, sediment type, etc.).

You can use the eMoF table to store these measurements by linking them to the correspondent EventID, the same way you did with the SamplingProtocol.

You can use the Vocabulary search (or other search tools provided) to look for the appropriate P01 BODC Parameter Usage Vocabulary concept/parameter. For example:

- Temperature of the water body: search for "temperature%water%body"
- Salinity of the water body with CTD: search for "salinity%CTD"
- Measurements related to lithology (sediment characteristics): "lithology"

Dealing with measurementTypes: Sampling facts

Searching your concept

Finally, we also need to populate the measurementTypeID field for all the records in the eMoF table that refer to the sampling method.

The dataset already contains information on the sampling method. We know that two types of sampling mechanisms were used:

- Gear: otter-trawl Maireta system (OTMS); The OTMS is a 1-warp benthic otter-trawl designed to work seamlessly on high depth grounds: its stretch mesh size at the cod-end is 40mm, with an outer cover of 12mm, to allow retrieval of small-sized fractions of megafauna. The net total length is 25m, with an horizontal opening of 12.7m and a vertical opening of 1.4m. Trawls were conducted at 2.6 to 2.8 knots.
- Gear: Agassiz dredge; The Agassiz dredge had a 2.5 m horizontal opening and 1.2 m vertical opening, a net mesh size of 12 mm, and was trawled at 2.0 knots.

How do we capture all this information in the eMoF table? Let's have a look at all the parameters that can be extracted from those sentences:

measurementType	**measurementValue&Unit**	
Gear/device type	Otter-Trawl Maireta System (OTMS)	Agassiz dredge
Mesh size	12mm	12mm
Net horizontal opening	12.7m	2.5m
Net vertical opening	1.4m	1.2m
Trawling speed	2.7knots (average)	2knots

Can you find the appropriate P01 parameters for each of the items in the first column?

Pro tip: you can use the Facet search to look for keywords (e.g. mesh size)

Populating measurementTypeID

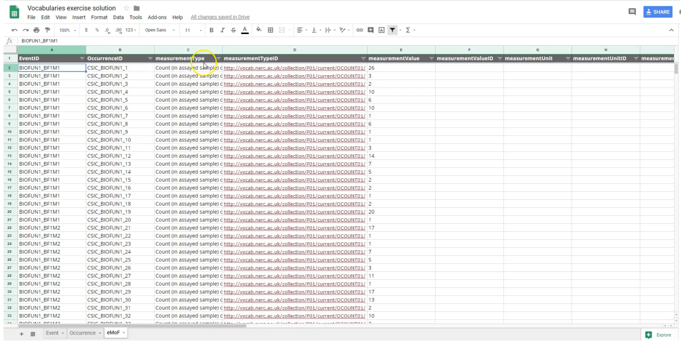
After looking for all the different concepts, our table would look like:

measurementType	**Label**	**Concept URI**	**measurementValue&Unit**	
Gear/device	Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)	Agassiz dredge
Mesh size	Sampling net mesh size	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100015/	12mm	12mm
Net horizontal opening	Sampling device aperture length	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100014/	12.7m	2.5m
Net vertical opening	Sampling device aperture width	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100013/	1.4m	1.2m

Trawling speed	Speed of measurement platform relative to ground surface {speed over ground}	http://vocab.nerc.ac.uk/collection/P01/current/APSAZZ01/	2.7knots (average)	2knots

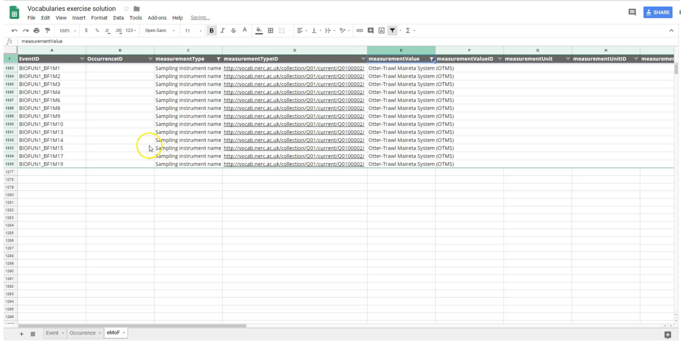
The steps we need to follow are shown in the following videos:

Sampling instrument name:



	EventID	OccurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID	measuremen
2	BIOFUN1_BF1M1	CSIC_BIOFUN1_1	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	26				
3	BIOFUN1_BF1M1	CSIC_BIOFUN1_2	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	3				
4	BIOFUN1_BF1M1	CSIC_BIOFUN1_3	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	2				
5	BIOFUN1_BF1M1	CSIC_BIOFUN1_4	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	10				
6	BIOFUN1_BF1M1	CSIC_BIOFUN1_5	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	6				
7	BIOFUN1_BF1M1	CSIC_BIOFUN1_6	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	10				
8	BIOFUN1_BF1M1	CSIC_BIOFUN1_7	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
9	BIOFUN1_BF1M1	CSIC_BIOFUN1_8	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	6				
10	BIOFUN1_BF1M1	CSIC_BIOFUN1_9	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
11	BIOFUN1_BF1M1	CSIC_BIOFUN1_10	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
12	BIOFUN1_BF1M1	CSIC_BIOFUN1_11	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	3				
13	BIOFUN1_BF1M1	CSIC_BIOFUN1_12	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	14				
14	BIOFUN1_BF1M1	CSIC_BIOFUN1_13	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	7				
15	BIOFUN1_BF1M1	CSIC_BIOFUN1_14	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	5				
16	BIOFUN1_BF1M1	CSIC_BIOFUN1_15	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	2				
17	BIOFUN1_BF1M1	CSIC_BIOFUN1_16	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	2				
18	BIOFUN1_BF1M1	CSIC_BIOFUN1_17	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
19	BIOFUN1_BF1M1	CSIC_BIOFUN1_18	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	2				
20	BIOFUN1_BF1M1	CSIC_BIOFUN1_19	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	20				
21	BIOFUN1_BF1M1	CSIC_BIOFUN1_20	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
22	BIOFUN1_BF1M2	CSIC_BIOFUN1_21	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	17				
23	BIOFUN1_BF1M2	CSIC_BIOFUN1_22	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
24	BIOFUN1_BF1M2	CSIC_BIOFUN1_23	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
25	BIOFUN1_BF1M2	CSIC_BIOFUN1_24	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	7				
26	BIOFUN1_BF1M2	CSIC_BIOFUN1_25	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	5				
27	BIOFUN1_BF1M2	CSIC_BIOFUN1_26	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	3				
28	BIOFUN1_BF1M2	CSIC_BIOFUN1_27	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	11				
29	BIOFUN1_BF1M2	CSIC_BIOFUN1_28	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	1				
30	BIOFUN1_BF1M2	CSIC_BIOFUN1_29	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	17				
31	BIOFUN1_BF1M2	CSIC_BIOFUN1_30	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	13				
32	BIOFUN1_BF1M2	CSIC_BIOFUN1_31	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	2				
33	BIOFUN1_BF1M2	CSIC_BIOFUN1_32	Count (in assayed sample) c	http://vocab.nerc.ac.uk/collection/P01/current/OCOUNT01/	10				

Sampling device aperture length:



	EventID	OccurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	measurementUnitID	measuremen
1223	BIOFUN1_BF1M1		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1224	BIOFUN1_BF1M2		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1225	BIOFUN1_BF1M3		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1226	BIOFUN1_BF1M4		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1227	BIOFUN1_BF1M6		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1228	BIOFUN1_BF1M8		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1229	BIOFUN1_BF1M9		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1230	BIOFUN1_BF1M10		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1231	BIOFUN1_BF1M13		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1232	BIOFUN1_BF1M14		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1233	BIOFUN1_BF1M15		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1234	BIOFUN1_BF1M17		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				
1235	BIOFUN1_BF1M19		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)				

Repeat these steps for the rest of the measurements.

Dealing with measurementValues

Searching your concept

You have now standardised all the measurement types, using controlled vocabularies from the P01 collection to populate the measurementTypeID field.

In some cases, it is also possible to give a controlled vocabulary for the measurement values. In other cases, it is not. For example, in your dataset you have numeric values for the abundances, biomass, etc. It does not make sense to have controlled vocabularies for the infinite possibilities offered by rational numbers (0.0005063291, 0.0000253165, etc.).

Typically, you can provide controlled vocabularies for other measurement or fact such as gender/sex, life/development stage, or the sampling devices. Below is again the list with the most common collections that you can look at in order to find vocabularies to standardise the measurement values:

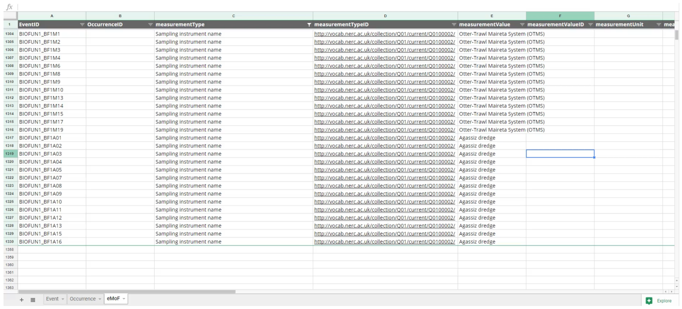
- S11- Biological entity life stage terms
- S10- BODC parameter semantic model biological entity gender terms
- L05- SeaDataNet device categories
- L22- SeaVoX Device Catalogue
- C17- ICES Platform Codes

Pro Tip: You can look for specific concepts within a collection by adding the collection code at the end of the URL of the NERC vocabulary search tool. For example: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/L22/

As it was mentioned before, in some cases, the BODC concept you are looking for is not yet part of the collection. If that is the case, contact EMODnet Biology.

Populating measurementValueID

You have now found a controlled vocabulary for the Aggasiz dredge. You will use it to populate the measurementValueID in the eMoF table. The only thing to do, is to paste the URI (http://vocab.nerc.ac.uk/collection/L22/current/TOOL1252/) into the measurementValueID field for all the records where measurementValue = Agassiz dredge.



	A	B	C	D	E	F	G	
1	EventID	OccurrenceID	measurementType	measurementTypeID	measurementValue	measurementValueID	measurementUnit	mea
1304	BIOFUN1_BF1M1		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1305	BIOFUN1_BF1M2		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1306	BIOFUN1_BF1M3		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1307	BIOFUN1_BF1M4		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1308	BIOFUN1_BF1M6		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1309	BIOFUN1_BF1M8		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1310	BIOFUN1_BF1M9		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1311	BIOFUN1_BF1M10		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1312	BIOFUN1_BF1M13		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1313	BIOFUN1_BF1M14		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1314	BIOFUN1_BF1M15		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1315	BIOFUN1_BF1M17		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1316	BIOFUN1_BF1M19		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Otter-Trawl Maireta System (OTMS)			
1317	BIOFUN1_BF1A01		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1318	BIOFUN1_BF1A02		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1319	BIOFUN1_BF1A03		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1320	BIOFUN1_BF1A04		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1321	BIOFUN1_BF1A05		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1322	BIOFUN1_BF1A07		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1323	BIOFUN1_BF1A08		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1324	BIOFUN1_BF1A09		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1325	BIOFUN1_BF1A10		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1326	BIOFUN1_BF1A11		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1327	BIOFUN1_BF1A12		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1328	BIOFUN1_BF1A13		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1329	BIOFUN1_BF1A15		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1330	BIOFUN1_BF1A16		Sampling instrument name	http://vocab.nerc.ac.uk/collection/Q01/current/Q0100002/	Agassiz dredge			
1358								
1359								
1360								
1361								
1362								
1363								

Note: There is no an adequate controlled vocabulary for Otter-Trawl Maireta System (OTMS). This means you cannot fill in measurementValueID for these records.

Dealing with measurementUnits

Searching for your concept

The last thing you need to do is to search for the standard vocabularies of the units of your measurements or facts. The collection where you need to look at is P06 (BODC units): http://vocab.nerc.ac.uk/collection/P06/current/

To search for a unit within this collection, remember you can use: https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/P06/

Try and look for the appropriate vocabularies for the units in your dataset:

- kg/km^2
- N/km^2
- mm
- m
- knots

Populating measurementUnitID

With this last step, your dataset will be ready to publish! Paste the controlled vocabulary that you have found in the correspondent measurementUnitID fields.

As you may have noticed, sometimes a suitable vocab term for your unit will not exist in P06, in these cases you can either contact EMODnet Biology to liaise with BODC in order to add a new vocabulary for this term, or you can convert your measurements to an unit that exists (remember that BODC uses SI units - find more about SI units here -)

measurements and facts without units

Units are necessary to properly understand the values measured. However, in some occasions you might have some measurements or facts that require no units to be fully understood, facts where the units are not applicable, dimensionless quantities, or units that have been deliberately not specified such as in arbitrary units.

As you can see if you have clicked in the previous links, controlled vocabulary terms exist to account for these cases and are therefore recommended to be used when corresponds. BODC has also created a vocabulary term for not known units that is recommended to be used if we are certain that the units exist but we just do not know them. This information can be used when querying data and in Quality Control procedures.

Why does data archaeology matter?

Although it considers observations far in the past, investing time in making these archaeological data available through online systems is of very high value.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Why does data archaeology matter?
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:29 PM

Table of contents

Why are historical data important?

Why are historical data important?

Historical or 'archaeological' data are important for many reasons...

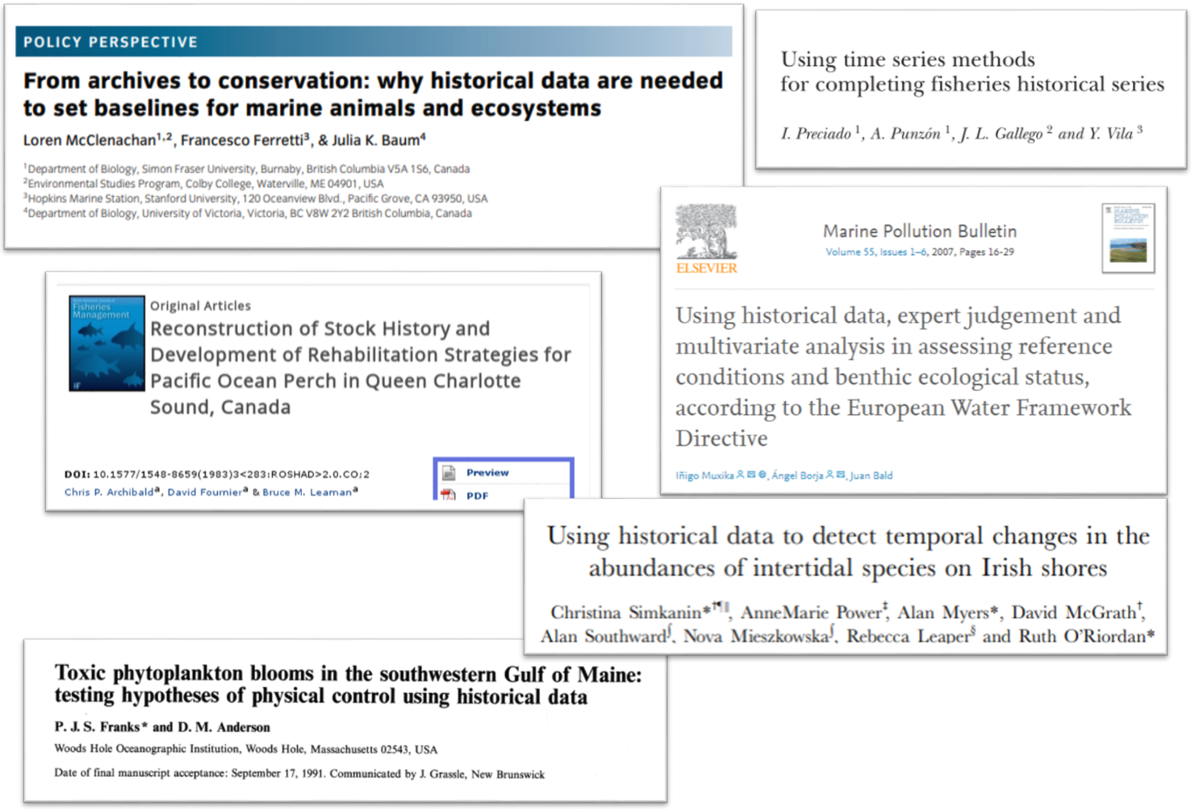
First of all, they provide the historical context for present observations, thereby facilitating the process of setting reference conditions for monitoring and management.

Linked to this, historical data can be used for reconstructing and modelling past conditions. They also have high value in predicting future trends and shifts in species distribution range, regional species extinctions, biological invasions, and consequences of human activities for the environment and biodiversity. This is extremely important for regional European Seas and adjacent marine regions, which are vulnerable to an ever-increasing number of human activities and pressures.

Historical datasets often contain descriptions of new species that are important for taxonomy as the first description of a species has legal priority for the name of this species.

Permanent loss of data has fatal consequences, simply because it is impossible to retrieve those data, which were collected from a certain area over a certain period. Consequently, loss of data equals to loss of unique resources and ultimately to the loss of our natural wealth.

The European Commission has always put forward the principle "Collect once, use many times", referring to all data ever collected. By reusing historical data in a comparing perspective, scientists can gain a lot of new insights...



POLICY PERSPECTIVE

From archives to conservation: why historical data are needed to set baselines for marine animals and ecosystems

Loren McClenachan[1,2], Francesco Ferretti[3], & Julia K. Baum[4]

[1] Department of Biology, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada
[2] Environmental Studies Program, Colby College, Waterville, ME 04901, USA
[3] Hopkins Marine Station, Stanford University, 120 Oceanview Blvd., Pacific Grove, CA 93950, USA
[4] Department of Biology, University of Victoria, Victoria, BC V8W 2Y2 British Columbia, Canada

Using time series methods
for completing fisheries historical series

I. Preciado [1], A. Punzón [1], J. L. Gallego [2] and Y. Vila [3]

Original Articles

Reconstruction of Stock History and Development of Rehabilitation Strategies for Pacific Ocean Perch in Queen Charlotte Sound, Canada

DOI: 10.1577/1548-8659(1983)3<283:ROSHAD>2.0.CO;2

Chris P. Archibald[a], David Fournier[a] & Bruce M. Leaman[a]

Marine Pollution Bulletin
Volume 55, Issues 1–6, 2007, Pages 16-29

Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive

Iñigo Muxika, Ángel Borja, Juan Bald

Using historical data to detect temporal changes in the abundances of intertidal species on Irish shores

Christina Simkanin[*¶‖], AnneMarie Power[‡], Alan Myers[*], David McGrath[†], Alan Southward[ʃ], Nova Mieszkowska[ʃ], Rebecca Leaper[§] and Ruth O'Riordan[*]

Toxic phytoplankton blooms in the southwestern Gulf of Maine: testing hypotheses of physical control using historical data

P. J. S. Franks* and D. M. Anderson

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

Date of final manuscript acceptance: September 17, 1991. Communicated by J. Grassle, New Brunswick

Difficulties & challenges in data archaeology

This chapter gives an overview of the kind of challenges you can encounter when working with archaeological data, and how you can overcome them.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Difficulties & challenges in data archaeology
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:29 PM

Table of contents

What can you expect?

Some examples

What can you expect?

Several things can make the work with archaeological data a real challenge. This section lists the most common ones.

Historical publications can lack standardisation among volumes, or even within the same publication. Well-known examples are e.g. authors who do not follow a specific format for presenting their results. How results are presented can vary from e.g. species lists to documenting (or not documenting) sampling dates and depths, while others also record individual species counts.

The information in these publications is often unstructured, hidden in free text, not annotated with metadata and thus not easily retrievable. In some cases, information from tables can be repeated in the text with small differences, making it very hard to understand which part of the given information is actually correct.

Different volumes of the same expedition can also be published in different languages, depending on the taxonomic expert who analysed the material and authored the publication. As a data manager, it can have its advantages to be a polyglot!

In many cases, coordinates can be missing, and only a location name gives an indication on where the samples have been collected. When coordinates are available, one can still stumble upon the problem of missing or inaccurate geo-referencing, e.g. when sample locations of marine species appear on land, e.g. in the center of an island. In some historical publications, coordinates may refer to the Paris and not the Greenwich meridian line, without a proper indication of this.

Historical publications often refer to

- uncommon/non-SI units of measurements which require conversion to international system of units
- old toponyms and political boundaries
- findings from other expeditions for which limited information is available

And they often include:

- taxonomic inconsistencies (e.g. outdated synonyms and old classification schemes) which require updating
- misspellings of taxa and locations
- ambiguous symbols that can be interpreted in different ways by different people

The solutions to these 'gaps' in the data can be addressed through several procedures that are explained throughout this course.

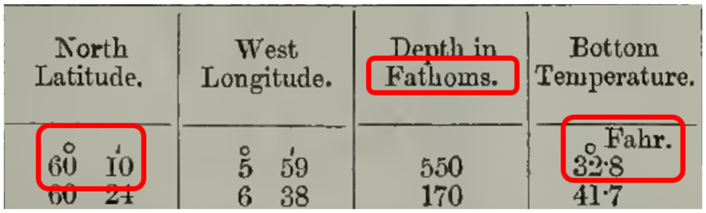
Two very common procedures in archaeology data management are:

1. Assignment of coordinates: when the latitude-longitude is not documented, it is still possible to pinpoint these - with a level of uncertainty - by using a Geographic Information System (GIS). If maps are available, these can be imported in GIS, and coordinates can be assigned. When specific places are identified, Marine Regions can be consulted to retrieve its coordinates.
2. In several cases new “artificial” sampling events should be created for samples (each sample is assumed to represent a unique combination of position, date, time, gear, length of wire and haul duration) due to inconsistencies between different expedition volumes and ambiguities which do not allow for a clear assignment to well-described sampling events.

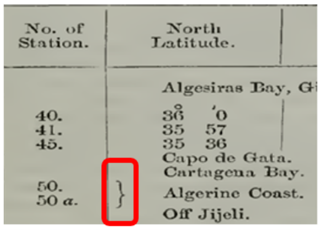
Some examples

The following are examples of some very frequent 'shortcomings' that one can encounter in archaeological data:

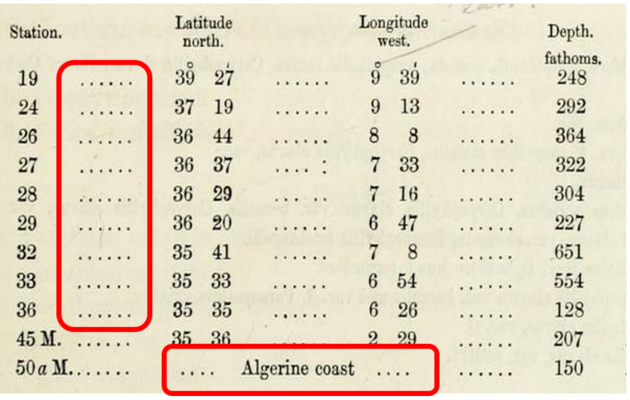
- Uncommon/non-SI units of measurement (e.g. 1 fathom ≈ 1.8 m)
- Interpretation of ambiguous symbols (e.g. asterisks vs. ditto marks vs. no-data)
- Missing information (e.g. general areas instead of coordinates)
- ...



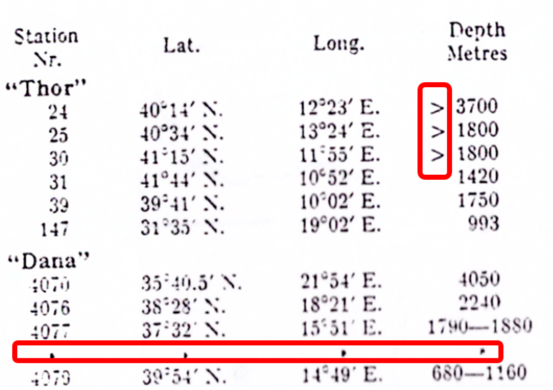
North Latitude.	West Longitude.	Depth in Fathoms.	Bottom Temperature.
° '	° '		Fahr.
60 10	5 59	550	32°·8
60 24	6 38	170	41·7



No. of Station.	North Latitude.
	Algesiras Bay, Gi
40.	36° 0'
41.	35 57
45.	35 36
	Capo de Gata.
	Cartagena Bay.
50. 50 *a*.	} Algerine Coast.
	Off Jijeli.



Station.		Latitude north.		Longitude west.		Depth. fathoms.
19	......	39° 27'	......	9° 39'	......	248
24	......	37 19	......	9 13	......	292
26	......	36 44	......	8 8	......	364
27	......	36 37	.... :	7 33	......	322
28	......	36 29	.... ..	7 16	......	304
29	......	36 20	......	6 47	......	227
32	......	35 41	......	7 8	......	651
33	......	35 33	......	6 54	......	554
36	......	35 35	......	6 26	......	128
45 M.	......	35 36	......	2 29	......	207
50 *a* M.	......	.... Algerine coast			......	150



Station Nr.	Lat.	Long.	Depth Metres
"Thor"			
24	40°14' N.	12°23' E.	> 3700
25	40°34' N.	13°24' E.	> 1800
30	41°15' N.	11°55' E.	> 1800
31	41°44' N.	10°52' E.	1420
39	39°41' N.	10°02' E.	1750
147	31°35' N.	19°02' E.	993
"Dana"			
4070	35°40.5' N.	21°54' E.	4050
4076	38°28' N.	18°21' E.	2240
4077	37°32' N.	15°51' E.	1790—1880
.	.	.	.
4079	39°54' N.	14°49' E.	680—1160

Digitisation workflow

Which steps need to be taken to make archaeological data available through modern data systems?

And where can you 'discover' archaeological data?

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Digitisation workflow
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:30 PM

Table of contents

Step by step: actual digitisation

Step by step: corrections

Where to find archaeological data?

An example: Historical oceanographic expeditions

Digitization workflow

Step by step: actual digitisation

Digitisation of archaeological data is not a quick process. It takes multiple steps and reiteration of steps to get to a high-quality digital dataset that can be integrated into larger initiatives.

It all starts with reading and comprehending the publication - or individual volumes of an expedition series - in order to overcome difficulties originating from the heterogeneity in the format and the content among the historical publications.

If you are digitising a series of volumes from a specific expedition, digitising the introductory volume (or logbook) is a good starting point. After that, each volume can be tackled independently, as distinct volumes usually concern different taxa or regions.

It is good practice to create an individual spreadsheet for each volume, and populate it with data - including e.g. species occurrences, abundances - and associated metadata.

You should not only digitise tables and appendices, but you should also check for complementary information scattered in the main text. This can vary from sampling methods over used sampling gear and any other observations, e.g. on general weather conditions or the state of the sea.

Step by step: corrections

Once information and data has been transferred from paper to a digital environment, some first corrections can be made, starting with the correction of obvious typographic errors.

In the process of checking the content, the introductory volume of the expedition should be used as a reference dataset. Based on this reference, unique sample IDs can be assigned throughout the different volumes of the expedition report. All information on location, date, time and gear type should be cross-referenced with this introductory volume as well.

Missing data in a dataset - e.g. coordinates, sampling gear and sampling time - could be either derived from the introductory of other volumes of the same expedition, whenever possible.

Very important is that species names and sampling location names in the digitised datasets should be kept exactly the same as in the original publication. Afterwards, all scientific names should be crosschecked and taxonomically updated using the Taxon Match tool of the World Register of Marine Species (WoRMS). Accepted names and their corresponding LSIDs should be listed in a separate column. It is of the utmost importance that the original name is kept, so it is always possible to retrace this name and its linked information.

Once all this is done, a second round of quality control can be undertaken on the integrated datasets. In cases of inconsistencies of information between the individual datasets and that of the introductory volume, it could be assumed that the information in the introductory volume is the correct one, keeping a note in the remarks field of the dataset.

All other aspects on how to format and quality control a dataset are similar to other - non archaeological - datasets and are described throughout the other chapters of this training course.

Where to find archaeological data?

Potential data sources:

- Institutional libraries
    - old publications
    - books
    - expedition logbooks
    - project reports
    - grey literature sources
- Online data sources
    - Biodiversity Heritage Library (BHL)
    - Reference module of the World Register of Marine Species

An example: Historical oceanographic expeditions

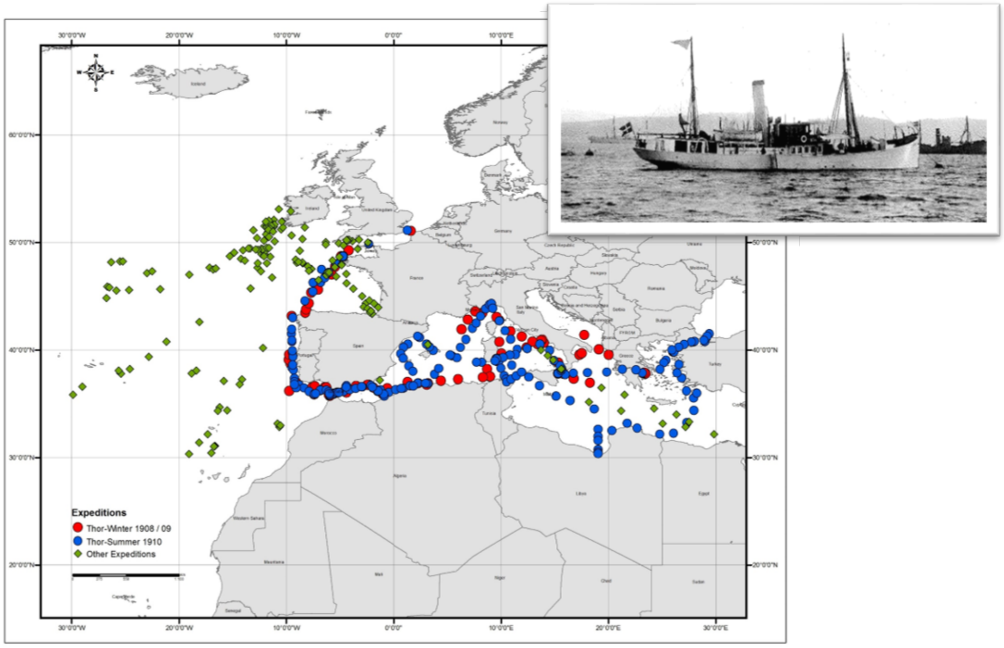
At the beginning of the 20th century, the importance of recording marine biodiversity was already recognized. Numerous expeditions had been organized with the aim of investigating “local fauna and flora” in various areas of the world.

During these scientific expeditions, local biodiversity of various taxonomic groups was collected, recorded and the outcome was published in many scientific volumes.

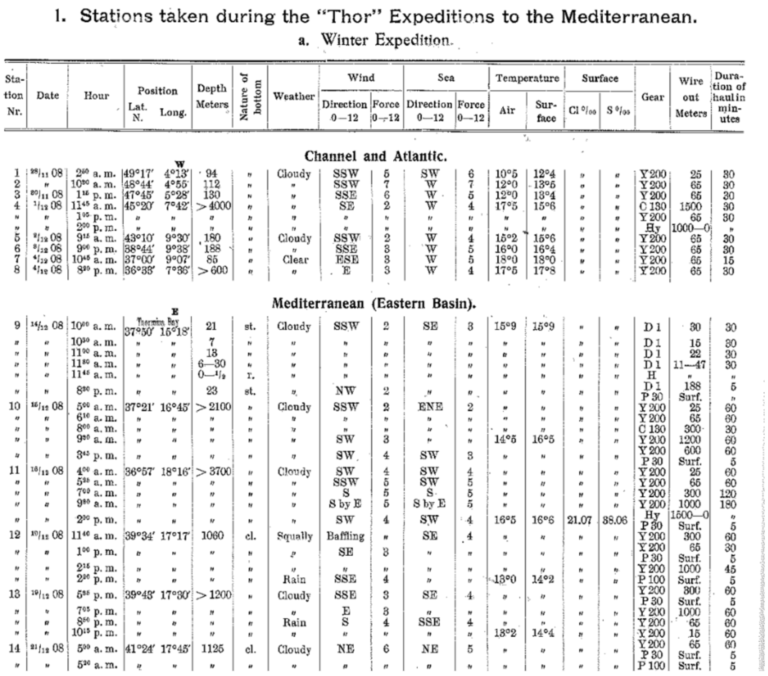
In most cases, the aims, sampling stations and metadata (e.g. coordinates, depths and gear type) were published in an introductory volume.

After sampling, all collected specimens were preserved and sent to several experts for taxonomic identification. Each expert was responsible for the publication of a different volume, which was often published in different languages and results were presented in different format and style.

The images below illustrate the variety in what an expedition report can look like, and the challenges a data digitizer can face.



Map of stations sampled by “Thor” and other vessels during the core expeditions of 1908-1909 and 1910 and additional expeditions from 1905-1906 and 1911-1912 (stations listed in the introductory table in Schmidt 1912). From Mavraki et al. 2016 BDJ.

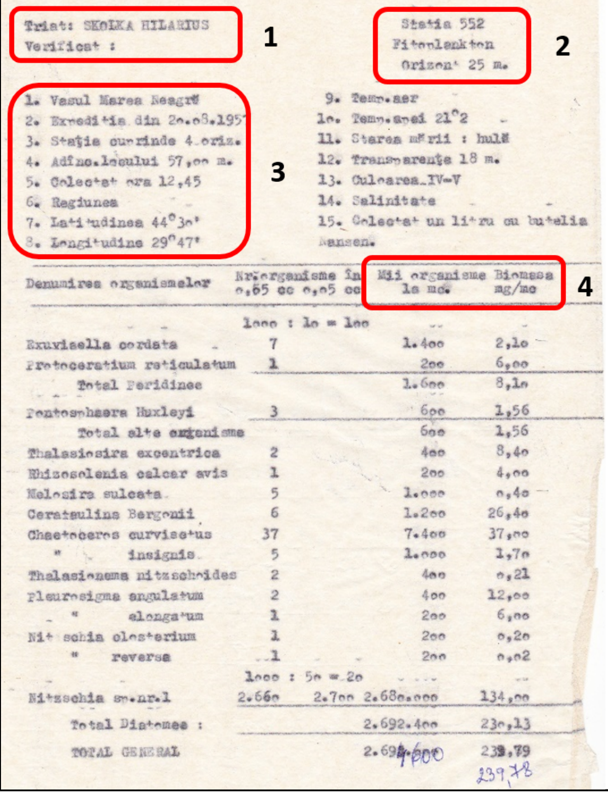


1. Stations taken during the "Thor" Expeditions to the Mediterranean.

a. Winter Expedition.

Station Nr.	Date	Hour	Position Lat. N.	Position Long.	Depth Meters	Nature of bottom	Weather	Wind Direction 0—12	Wind Force 0—12	Sea Direction 0—12	Sea Force 0—12	Temperature Air	Temperature Surface	Surface Cl ‰	Surface S ‰	Gear	Wire out Meters	Duration of haul in minutes
							Channel and Atlantic.											
1	28/11 08	2^{50} a. m.	49°17′	W 4°13′	94	″	Cloudy	SSW	5	SW	6	10°5	12°4	″	″	Y 200	25	30
2	″	10^{30} a. m.	48°44′	4°55′	112	″	″	SSW	7	W	7	12°0	13°5	″	″	Y 200	65	30
3	20/11 08	1^{15} p. m.	47°45′	5°28′	130	″	″	SSE	6	W	5	12°0	13°4	″	″	Y 200	65	30
4	1/12 08	11^{45} a. m.	45°20′	7°42′	> 4000	″	″	SE	2	W	4	17°5	15°6	″	″	C 130	1500	30
″	″	1^{05} p. m.	″	″	″	″	″	″	″	″	″	″	″	″	″	Y 200	65	30
″	″	2^{00} p. m.	″	″	″	″	″	″	″	″	″	″	″	″	″	Hy	1000—0	″
5	2/12 08	9^{15} a. m.	43°10′	9°30′	180	″	Cloudy	SSW	2	W	4	15°2	15°6	″	″	Y 200	65	30
6	3/12 08	9^{00} p. m.	38°44′	9°38′	188	″	″	SSE	3	W	5	16°0	16°4	″	″	Y 200	65	30
7	4/12 08	10^{45} a. m.	37°00′	9°07′	85	″	Clear	ESE	3	W	5	18°0	18°0	″	″	Y 200	65	15
8	4/12 08	8^{30} p. m.	36°33′	7°36′	> 600	″	″	E	3	W	4	17°5	17°8	″	″	Y 200	65	30
							Mediterranean (Eastern Basin).											
9	14/12 08	10^{00} a. m.	Taormina Bay 37°50′	E 15°18′	21	st.	Cloudy	SSW	2	SE	3	15°9	15°9	″	″	D 1	30	30
″	″	10^{30} a. m.	″	″	7	″	″	″	″	″	″	″	″	″	″	D 1	15	30
″	″	11^{00} a. m.	″	″	13	″	″	″	″	″	″	″	″	″	″	D 1	22	30
″	″	11^{30} a. m.	″	″	6—30	″	″	″	″	″	″	″	″	″	″	D 1	11—47	30
″	″	11^{45} a. m.	″	″	0—1/2	r.	″	″	″	″	″	″	″	″	″	H	″	″
″	″	8^{30} p. m.	″	″	23	st.	″	NW	2	″	″	″	″	″	″	D 1 P 30	188 Surf.	5 ″
10	16/12 08	5^{00} a. m.	37°21′	16°45′	> 2100	″	Cloudy	SSW	2	ENE	2	″	″	″	″	Y 200	25	60
″	″	6^{10} a. m.	″	″	″	″	″	″	″	″	″	″	″	″	″	Y 200	65	60
″	″	8^{00} a. m.	″	″	″	″	″	″	″	″	″	″	″	″	″	C 130	300	30
″	″	9^{30} a. m.	″	″	″	″	″	SW	3	″	″	14°5	16°5	″	″	Y 200	1200	60
″	″	3^{45} p. m.	″	″	″	″	″	SW	4	SW	3	″	″	″	″	Y 200 P 30	600 Surf.	60 5
11	18/12 08	4^{00} a. m.	36°57′	18°16′	> 3700	″	Cloudy	SW	4	SW	4	″	″	″	″	Y 200	25	60
″	″	5^{35} a. m.	″	″	″	″	″	SSW	5	SW	5	″	″	″	″	Y 200	65	60
″	″	7^{00} a. m.	″	″	″	″	″	S	5	S	5	″	″	″	″	Y 200	300	120
″	″	9^{30} a. m.	″	″	″	″	″	S by E	5	S by E	5	″	″	″	″	Y 200	1000	180
″	″	2^{00} p. m.	″	″	″	″	″	SW	4	SW	4	16°5	16°6	21.07	38.06	Hy P 30	1500—0 Surf.	″ 5
12	19/12 08	11^{40} a. m.	39°34′	17°17′	1060	cl.	Squally	Baffling	″	SE	4	″	″	″	″	Y 200	300	60
″	″	1^{00} p. m.	″	″	″	″	″	SE	3	″	″	″	″	″	″	Y 200 P 30	65 Surf.	30 5
″	″	2^{15} p. m.	″	″	″	″	″	″	″	″	″	″	″	″	″	Y 200	1000	45
″	″	2^{30} p. m.	″	″	″	″	Rain	SSE	4	″	″	13°0	14°2	″	″	P 100	Surf.	5
13	19/12 08	5^{35} p. m.	39°43′	17°30′	> 1200	″	Cloudy	SSE	3	SE	4	″	″	″	″	Y 200 P 30	300 Surf.	60 5
″	″	7^{05} p. m.	″	″	″	″	″	E	3	″	″	″	″	″	″	Y 200	1000	60
″	″	8^{50} p. m.	″	″	″	″	Rain	S	4	SSE	4	″	″	″	″	Y 200	65	60
″	″	10^{15} p. m.	″	″	″	″	″	″	″	″	″	13°2	14°4	″	″	Y 200	15	60
14	21/12 08	5^{00} a. m.	41°24′	17°45′	1125	cl.	Cloudy	NE	6	NE	5	″	″	″	″	Y 200 P 30	65 Surf.	60 5
″	″	5^{20} a. m.	″	″	″	″	″	″	″	″	″	″	″	″	″	P 100	Surf.	5

Original presentation of sampling metadata of the historical "Thor" Expedition to the Mediterranean Sea and adjacent waters (from Schmidt 1912)



Triat: SKOLKA HILARIUS
Verificat :

Statia 552
Fitoplankton
Orizont 25 m.

1. Vasul Marea Neagră
2. Expeditia din 2o.o8.195
3. Staţie cuprinde 4 oriz.
4. Adînc.locului 57,oo m.
5. Colectat ora 12,45
6. Regiunea
7. Latitudinea 44°3o'
8. Longitudine 29°47'

9. Temp.aer
1o. Temp.apei 21°2
11. Starea mării : hulă
12. Transparenţa 18 m.
13. Culoarea IV–V
14. Salinitate
15. Colectat un litru cu butelia Nansen.

Denumirea organismelor	Nr.organisme în o,65 cc	o,o5 cc	Mii organisme la mc.	Biomasa mg/mc
	1ooo : 1o = 1oo			
Exuviaella cordata	7		1.4oo	2,1o
Protoceratium reticulatum	1		2oo	6,oo
Total Peridinee			1.6oo	8,1o
Pontosphaera Huxleyi	3		6oo	1,56
Total alte organisme			6oo	1,56
Thalassiosira excentrica	2		4oo	8,4o
Rhizosolenia calcar avis	1		2oo	4,oo
Melosira sulcata	5		1.ooo	o,4o
Cerataulina Bergonii	6		1.2oo	26,4o
Chaetoceros curvisetus	37		7.4oo	37,oo
" insignis	5		1.ooo	1,7o
Thalassionema nitzschoides	2		4oo	o,21
Pleurosigma angulatum	2		4oo	12,oo
" elongatum	1		2oo	6,oo
Nitzschia closterium	1		2oo	o,2o
" reversa	1		2oo	o,o2
	1ooo : 5o = 2o			
Nitzschia sp.nr.1	2.66o	2.7oo	2.68o.ooo	134,oo
Total Diatomee :			2.692.4oo	23o,13
TOTAL GENERAL			2.694.600	239,79 239,78

Original data protocol

of phytoplankton sample from the Romanian Black Sea (1957)

1. Name of person who analysed the sample

2. Number of sample and layer sampled

3. General information about research vessel, expedition date, station depth, sampling hour and coordinates

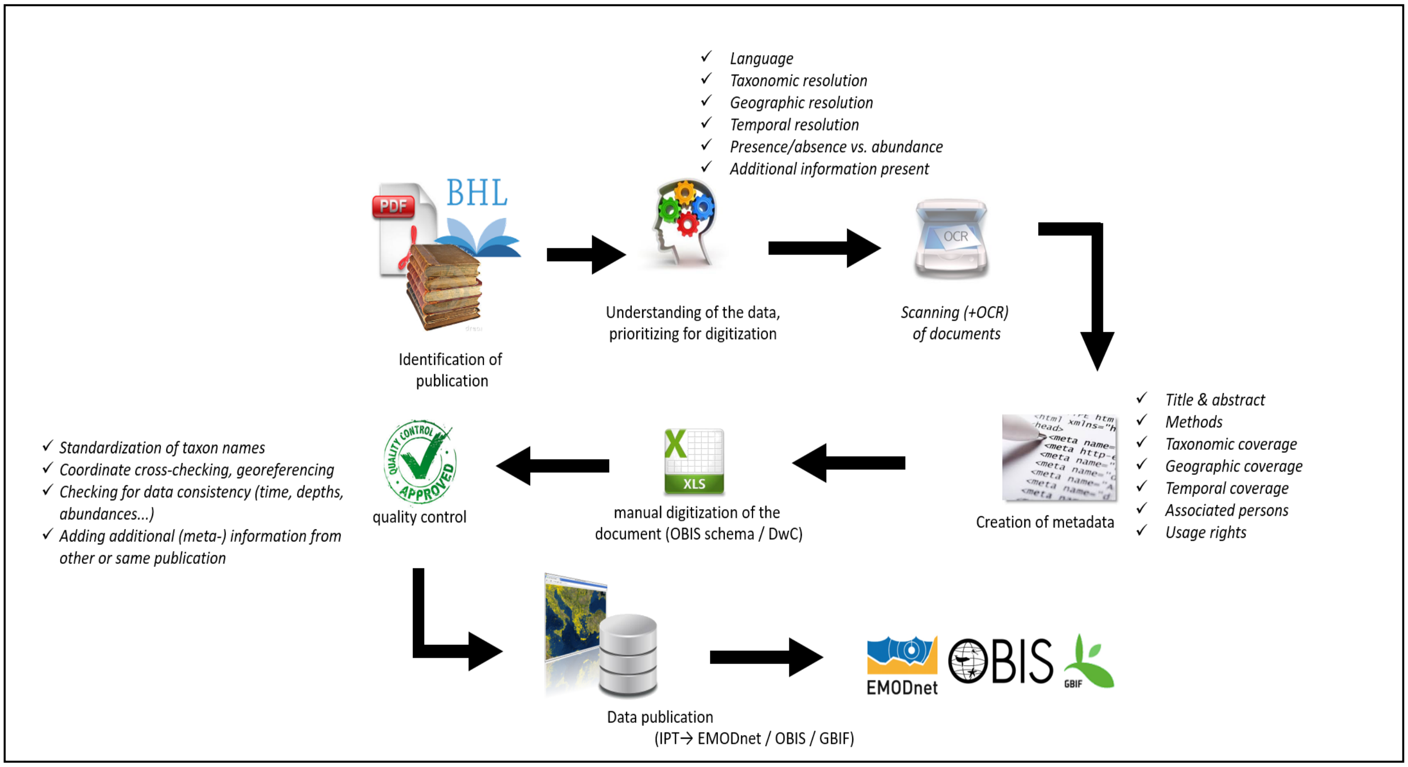
4. Quantitative data of abundance and biomass

Source: NIMRD biological data archive (provided by Laura Boicenco)

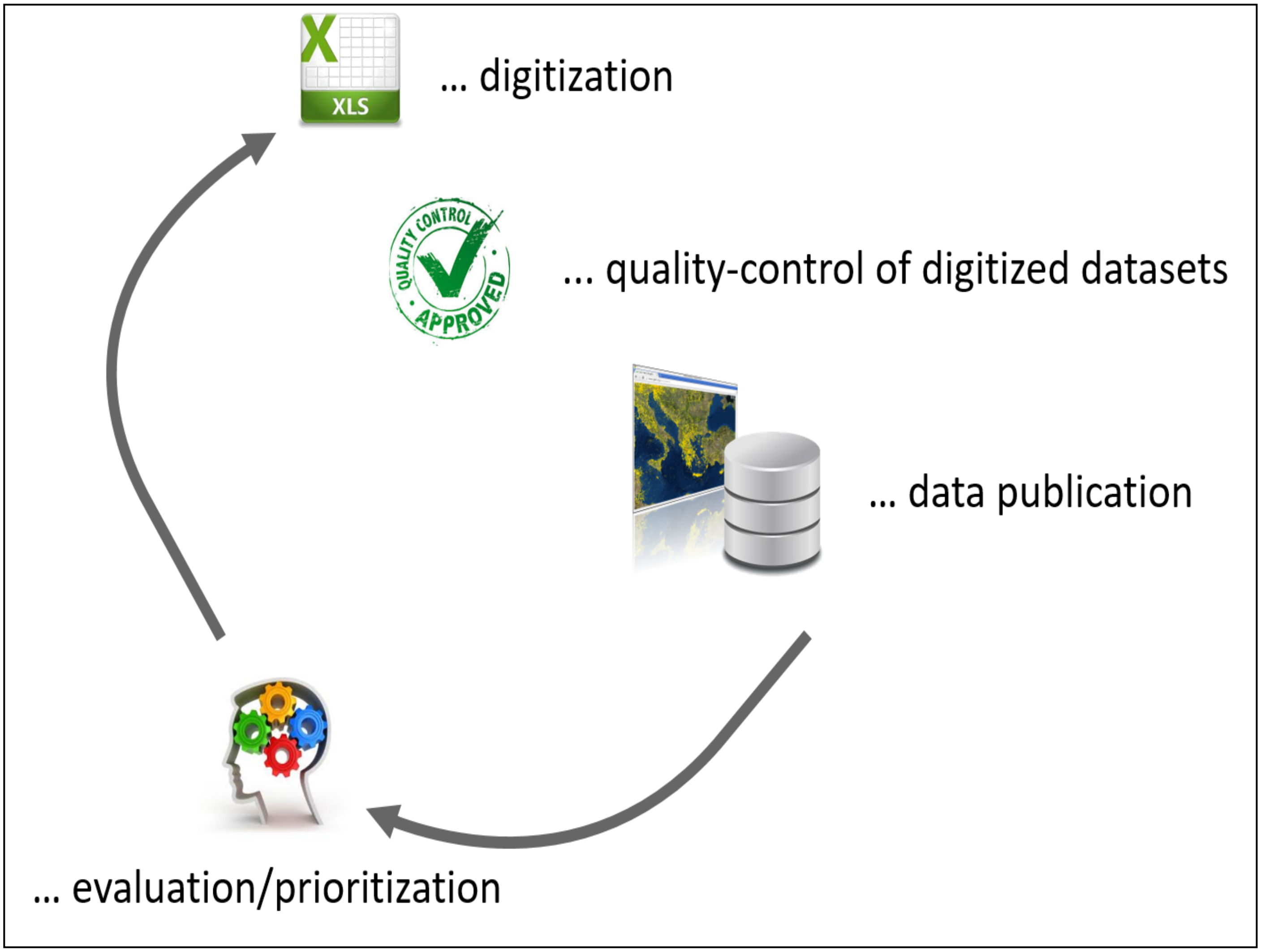
Digitization workflow

The digitization of data consists of many different steps, each step with its own peculiarities to take into account.

It is a continuous process, which can be time-consuming but in the end is a very valuable and worthwhile exercise.



BHL
PDF
Identification of publication
✓ Language
✓ Taxonomic resolution
✓ Geographic resolution
✓ Temporal resolution
✓ Presence/absence vs. abundance
✓ Additional information present
Understanding of the data, prioritizing for digitization
OCR
Scanning (+OCR) of documents
✓ Title & abstract
✓ Methods
✓ Taxonomic coverage
✓ Geographic coverage
✓ Temporal coverage
✓ Associated persons
✓ Usage rights
Creation of metadata
XLS
manual digitization of the document (OBIS schema / DwC)
QUALITY CONTROL APPROVED
quality control
✓ Standardization of taxon names
✓ Coordinate cross-checking, georeferencing
✓ Checking for data consistency (time, depths, abundances…)
✓ Adding additional (meta-) information from other or same publication
Data publication (IPT→ EMODnet / OBIS / GBIF)
EMODnet
OBIS
GBIF



... digitization

... quality-control of digitized datasets

... data publication

... evaluation/prioritization

In conclusion

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	In conclusion
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:30 PM

Table of contents

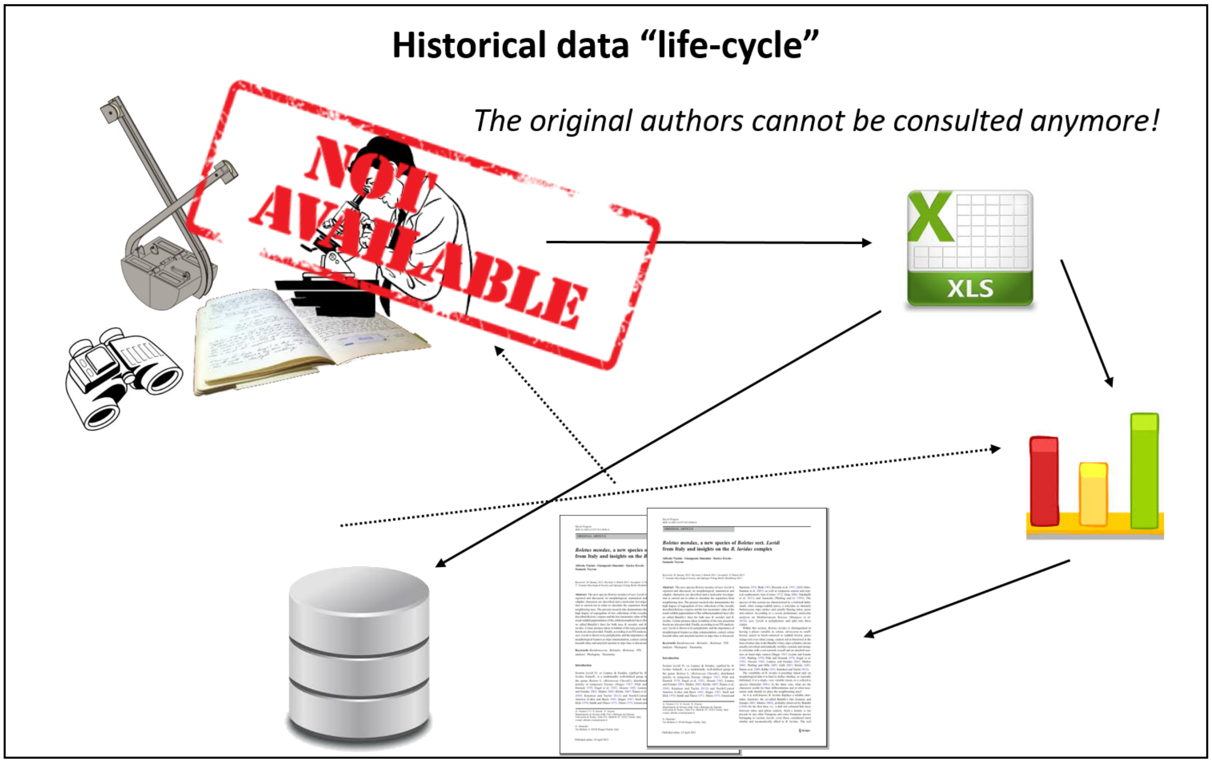
in conclusion

in conclusion

When legacy data are concerned, their integration into a harmonised, comparable format and their quality control is a special challenge.

Data might be inconsistent across different volumes of the same expedition, or even within the same publication, and obvious errors are identified during digitization. Often, information is missing, or spread across several publications.

As the original authors can no longer be consulted, some of these problems will remain unresolved indefinitely, and there is currently no standard strategy or “best practice” on how to deal with such issues.



Historical data “life-cycle”

The original authors cannot be consulted anymore!

For all the reasons discussed in this chapter, the extraction of archaeological information can be a very tedious and time-consuming effort.

Nevertheless, biodiversity legacy literature contains a wealth of information on the biosphere and can provide valuable insights into the past state of the world’s ecosystems. Consequently, loss of data equals to loss of unique resources and ultimately to the loss of our natural wealth.

How to publish the dataset through IPT?

You will learn how to use the Integrated Publishing Toolkit (IPT) to transform your dataset into a Darwin Core Archive file as well as how to install your own IPT instance and the advantages of administering one.

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	How to publish the dataset through IPT?
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:30 PM

Table of contents

Introduction

Metadata

How to create an IPT resource

Publish your dataset

Technical

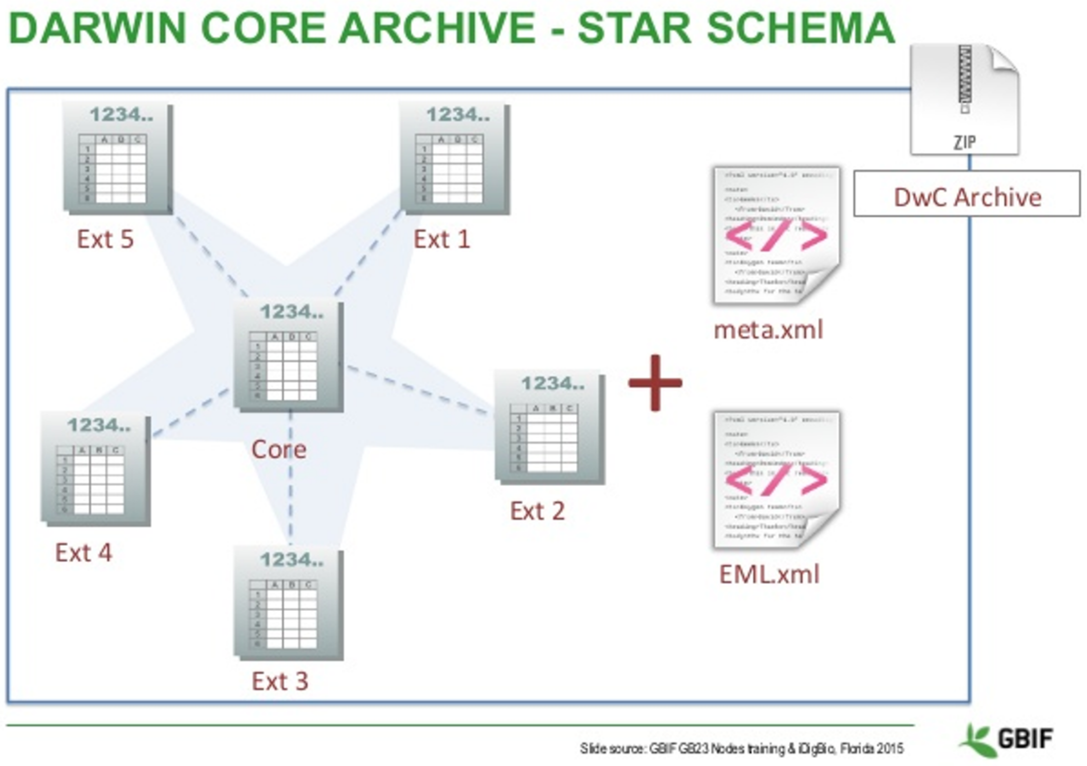
Introduction

In biological data management GBIF's Integrated Publishing Toolkit (IPT) has become the standard software for exchanging data. The IPT has been developed by GBIF to allow institutes to automatically send their data to GBIF. It has also been adopted by OBIS and EMODnet Biology as the preferred way to harvest data. We will discuss the dataflow in detail in the "Data harvest & DOIs" book of this course.

The IPT is a tool that creates DwC Archive (DwC-A) files which are self explanatory dataset resources and contain the following files:

- **eml.xml**: contains all dataset's metadata
- **meta.xml**: lists the different terms used in the data files and explains the relationship between them, such as which data file is the core, and which data files are extensions to the core.
- **event, occurrence, extendedmeasurementorfact**: these 3 data files (.csv or .txt) correspond to the 3 data tables mentioned in the “Data structure” book of this course.

Below you find a schematic representation of a DwC-A file. An example of a DwC-A can be downloaded here.


DARWIN CORE ARCHIVE - STAR SCHEMA
ZIP
DwC Archive
Ext 5
Ext 1
Core
Ext 2
Ext 4
Ext 3
meta.xml
EML.xml
+
Slide source: GBIF GB23 Nodes training & iDigBio, Florida 2015
GBIF

The IPT helps organising the DwC-A files and facilitates the exchange between different systems. It also displays the metadata for each resource (dataset) in a human readable manner and allows for the download of the DwC-A .zip file.

All the dataset in OBIS that flow from EurOBIS are available at the EurOBIS IPT. Take a moment to look at the resources in that page. Download a DwC-A file and get acquainted with IPT.

Metadata

To create an IPT resource you will need to add both metadata and data on IPT. In the “Metadata in EMODnet Biology” book of this course we discussed which metadata is required for EMODnet Biology; the IPT mandatory metadata is the same. Please, make sure you add the same metadata as provided for EMODnet Biology (pay special attention in using the same title!) and remember that metadata (and data) must not be modified without the consent of the data provider.

Below you can find again the most relevant metadata for EMODnet Biology. Read it through to refresh your memory

Title

Upon creating an IPT resource, IPT will require a *Shortname*. The *Shortname* is the identifier used as a parameter in the URL to access the IPT resource. This *Shortname* must be composed by alphanumerics, hyphens and/or underscores and as its very name suggest, the *Shortname* needs to be concise.

The *Title* of the dataset serves as an identifier and will play a major role in the discoverability of the dataset. Therefore, it should be as descriptive and complete as possible. EMODnet Biology recommends titles to contain information about the taxonomic, geographic and temporal coverage. In IPT, the Title section will be automatically populated with the *Shortname*, make sure to replace it with the actual *Title*.

As an example for http://ipt.vliz.be/eurobis/resource.do?r=deepsea_echinoidea:

- *Shortname* = “deepsea_echinoidea”
- *Title* = “Echinoidea distribution data from: Deep-sea fauna of European seas…”

Description

It should be the same as the Abstract metadata field provided to EMODnet Biology

People and Organizations

The Ecological Metadata Language (please see EML) has several possible roles/functions to describe a contact, creator, metadata provider and associated party.

Upon filling in the IPT resource metadata these roles will also be available. Have a look at the mapping between IMIS and IPT “People and Organizations” below:

Integrated Marine Information System (IMIS)	Integrated Publishing Toolkit (IPT)
Dataset contact	Resource Contact
Data creator	Resource Creators
Person providing the metadata	Metadata Providers

Note: *When adding affiliations for data creators, make sure to add the organisation they were working for when the data were being gathered and analysed. As is the case in the author's affiliation in literature publications, the data creator's affiliation* *refers to employment during the creation of the resource. As such the affiliation of a person added under contact may differ from the affiliation from the same person added under resource creator.*

Licence

For compatibility reasons, by default, IPT only allows you to choose from 3 Creative Commons licenses just like EMODnet Biology:

- CC-0 public domain: CC-0 is the preferred option identified by the OBIS steering group and GBIF. 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.
- CC-BY Attribution
- CC-BY-NC non-commercial

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).

The information given in this section can also help the data 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 area, 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 description field. Note that the IPT only allows one bounding box.

Taxonomic Coverage

This section can capture the taxonomic coverage in two ways:

- A description of the range of taxa that are addressed in the dataset. EMODnet Biology recommends adding only 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.
- An overview of all the involved taxa (not recommended, as all the taxa are already listed in the dataset).

Temporal Coverage

The temporal coverage will be a date range that can easily be documented. If it is a single date, the start and end dates will be the same. The information added here can be used as a quality check for the actual dates in the dataset. For ongoing projects contact EMODnet Biology.

Keywords

It is recommended to the add the same Keywords as provided to EMODnet Biology

Citations

The dataset citation allows users to properly cite the datasets in other publications or other uses of the data. *A dataset citation is different from the data source citation* (in case the data is digitised from a publication). Citations are structured as follows.

{resource.creators} ({dataset.pubDate}) {dataset.title}. {Publisher.title}. {dataset.DOI}

We will discuss the DOI and the Publisher parts in more detail in the "Data harvest & DOIs" book of this course.

Bibliographic Citations

This field is similar to the "Publications related to the dataset" metadata field provided to EMODnet Biology

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.

External Links

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

How to create an IPT resource

Once your data files have been formatted according to the EMODnet Biology (and EurOBIS) guidelines, they will need to be added to the IPT resource as well. Pro Tip: store your data files as tab delimited .txt before adding them to the IPT resource.

Please watch this demo video by GBIF to see how to create an IPT resource from scratch and how to add data and metadata to it.

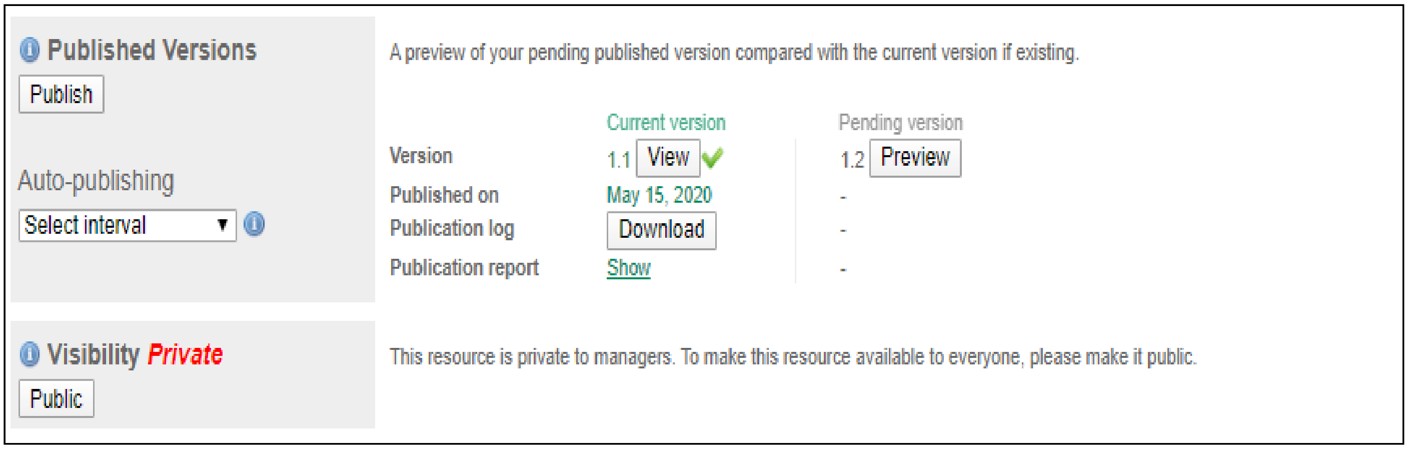
Please bear in mind that in the video example they use the Occurrence as the Core mapping while the OBIS-ENV data format followed by EMODnet Biology uses the Event as the Core mapping, using the Darwin Occurrence as an Extension table.

Publish your dataset

Once the data have been mapped as demonstrated in the previous video and the metadata have been filled in, the IPT resource need to be published.

In the “Overview” page of the IPT resource > Visibility > click on Public AND in Published Versions > click on “Publish” to commit your changes and generate your dataset as a DwC-A

<u>Attention</u>: The dataset is now available to everyone. If you want to make the dataset "Private", click on Visibility > Private AND click on “Publish” again to commit the changes.



A preview of your pending published version compared with the current version if existing.

	Current version	Pending version
Version	1.1 View ✓	1.2 Preview
Published on	May 15, 2020	-
Publication log	Download	-
Publication report	Show	-

Visibility *Private*

This resource is private to managers. To make this resource available to everyone, please make it public.

Your published dataset contains a static snapshot of your data. To update the data, you will need to add and map the updated data files, remove (or unmap) the old data files and “Publish” the changes.

Technical

Install the IPT

Now you may be thinking "the IPT is great, but how do I get one?"

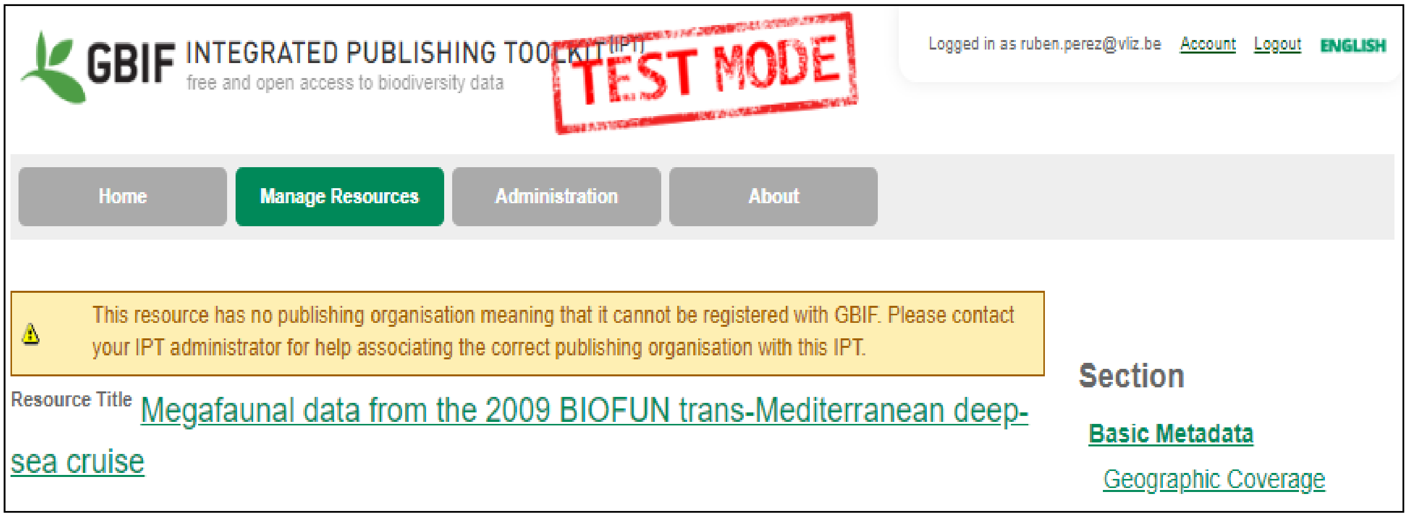
Well, you have the option of setting up an IPT on your own server or you could ask EMODnet Biology to set one up at the VLIZ server for you. The VLIZ IPT installation currently serves a number of OBIS nodes and EMODnet biology providers. VLIZ offers this service because - while creating an IPT is rather easy and it's well documented (https://github.com/gbif/ipt/wiki) - the IPT does require some maintenance as it needs to be kept up to date with the most recent versions.

- If you wish to have your own IPT installed at VLIZ or if you have any issues upon installing, please contact bio@EMODnet.eu.
- If you wish to host an IPT yourself, you can find documentation on how to do this at https://github.com/gbif/ipt/wiki/IPT2ManualNotes.wiki#install-the-ipt

Administer the IPT

When VLIZ sets up an IPT for you it's typically set up in test mode. An IPT in test mode has all the functionalities as an IPT in production mode with one key difference: you can not send data to GBIF.

You can see if an IPT is in test mode by 'test mode' stamp in the header as shown in the image below from http://ipt.iobis.org/training:


GBIF INTEGRATED PUBLISHING TOOLKIT (IPT)
free and open access to biodiversity data
TEST MODE
Logged in as ruben.perez@vliz.be
Account
Logout
ENGLISH
Home
Manage Resources
Administration
About

⚠ This resource has no publishing organisation meaning that it cannot be registered with GBIF. Please contact your IPT administrator for help associating the correct publishing organisation with this IPT.

Resource Title Megafaunal data from the 2009 BIOFUN trans-Mediterranean deep-sea cruise

Section
Basic Metadata
Geographic Coverage

IPTs in test mode usually display the warning message (yellow box) shown in the image above upon adding metadata. However the IPT is fully operational and can be harvested by EMODnet Biology and OBIS. Data flow to GBIF will be discussed in more detail in the "Data harvest & DOIs" book of this course.

As the administrator of your IPT you are able to:

- Add users who can add datasets
- Install the extensions required by EMODnet Biology. Make sure that the following 3 are installed and updated to the latest version:

- Darwin Core Event
- Darwin Core Occurrence
- Extended Measurement Or Facts

IPT datasets and Quality Control

You will learn how to Quality Control your dataset using different tools to identify errors and flaws

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	IPT datasets and Quality Control
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:31 PM

Table of contents

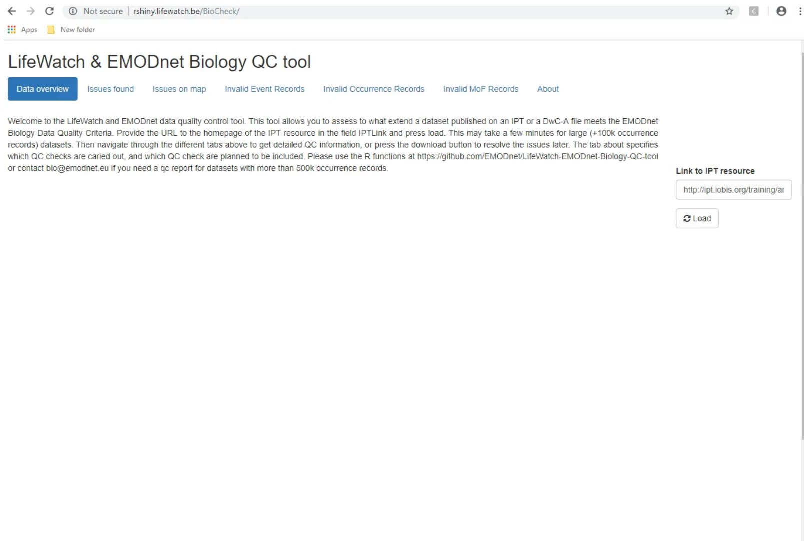
LifeWatch-EMODnet Biology QC tool

OBIS tools

LifeWatch-EMODnet Biology QC tool

As the dataset is now in an uniform format, you can run tools on it to verify if it meets the requirement. The LifeWatch-EMODnet Biology QC tool (BioCheck) is available online and you will be asked to use it on your version of the demo dataset as part of this course.

The video below explains the features of the LifeWatch-EMODnet Biology QC tool.



LifeWatch & EMODnet Biology QC tool

Welcome to the LifeWatch and EMODnet data quality control tool. This tool allows you to assess to what extend a dataset published on an IPT or a DwC-A file meets the EMODnet Biology Data Quality Criteria. Provide the URL to the homepage of the IPT resource in the field IPTLink and press load. This may take a few minutes for large (+100k occurrence records) datasets. Then navigate through the different tabs above to get detailed QC information, or press the download button to resolve the issues later. The tab about specifies which QC checks are caried out, and which QC check are planned to be included. Please use the R functions at https://github.com/EMODnet/LifeWatch-EMODnet-Biology-QC-tool or contact bio@emodnet.eu if you need a qc report for datasets with more than 500k occurrence records.

Link to IPT resource: http://ipt.iobis.org/training/ar

**Keep in mind that the IPT resource needs to be "Public" (find more in the "How to publish the dataset through IPT?" book of this course) to be read by the LifeWatch-EMODnet Biology QC tool*

The LifeWatch-EMODnet Biology QC tool runs the following checks:

- *Is the dataset integrity ok?Are all mandatory fields present in the dataset?*
    - *Do all eventID's in the occurrence extension refer to an Event Record?*
    - *Do all eventID's in the eMoF extension refer to an Event Record?*
    - *Do all occurrenceID's in the eMoF extension refer to an occurrence Record?*
    - *In case of a biometrical parameter, is the eventID in the eMoF table the same one as in the occurrence extension for the related occurrenceID?*
    - *Are there 'duplicate occurrences': is the same taxon listed twice within the same EventID without any difference in any of the biometric parameters?*
    - *Are there 'duplicate measurements': does the same measurement occur twice for the same occurrenceID or the same EventID?*
    - *Are all parentEventIDs linked to an existing event record?*
- *Are all mandatory fields filled out?*
- *Does eventDate follow the required format?*
- *Do the scientificNameIDs follow the required format?*
- *Are there coordinates located on land? (buffer of 3km is taken under consideration)*
- *Do the coordinates on land refer to marine taxa?*
- *Are there depths at the location deeper than the depths stored by EMODnet bathymetry (Europe) and GEBCO (the rest of the world)? (a margin of 150m is taken under consideration)*

- *Do all measurementtypes have a MeasurementTypeID?*
- *Do the measurementTypeIDs / measurementValueIDs refer to an existing term in the BODC vocabulary? (not all the BODC collections are used)*
- *Do all measurementValues that refer to facts have a measurementValueID?*
- *Are there records where measurementValue is NULL?*
- *Are there records that refer to biological measurement where measurementValue = 0 (and where occurrenceStatus is not absent)?*
- *Provide overview of the measurementTypes, their units, the min and max values and the pref labels, definitions and standard units associated to the MeasurementTypeID*
- *Provide an overview of the number of taxa per kingdom and class.*
- *List the non-matched taxa (including deleted and quarantined matches)*
- *Is there a sampling instrument present?*
- *Are there other sampling descriptors present?*
- *Plots of the coordinates and the distribution of the temporal cover are provided to allow for quick comparison with the metadata.*
- *Option to display and download the OBIS tree structure: https://github.com/iobis/obistools/#dataset-structure*

Although the LifeWatch-EMODnet Biology QC tool helps finding many of the issues within a dataset, the data manager still has to do some extra checks. A number of additional checks and features are still planned to be implemented by the EMODnet Biology data management team to aid the data managers:

- *Does the occurrenceStatus contain standardised terms?*
- *Does the basisOfRecord contain standardised terms?*
- *Does the scientificName contain anything else than the scientificName?*
- *Have the scientificNameIDs been correctly assigned?*
- *Are there non-marine / non-brackish taxa found at sea?*
- *Have the units been provided for measurementTypes that require an unit to be correctly interpreted?*
- *Is the unit provided the same base unit as specified by the measurementTypeID?*
- *Explain what is meant by a "potentially duplicate record"*
- *Is the datasetName field different than the title of the dataset?*
- *...*

For error reports or feature requests, please contact bio@emodnet.eu

Upon submission to EMODnet Biology, the data management team will run a script that will check the DwC-A file and carry out the previously mentioned checks. (If you did all steps explained in the previous sections - especially in the "Technical metadata and data entity integrity" book of this course - there is nothing to worry about).

OBIS tools

OBIS tools R-package

As EMODnet Biology follows the same format as OBIS, all the tools developed by OBIS will be useful to either verify the quality of a dataset or to correct any mistakes.

The functions specified on https://github.com/iobis/obistools can be used on these data.

Functions which can be useful for data processing are:

- Taxon matching
- Check required fields
- Plot points on a map
- Identify points on a map
- Check points on land
- Check depth
- Check outliers
- Check eventID and parentEventID
- Check eventID in an extension
- Flatten event records
- Flatten occurrence and event records
- Calculate centroid and radius for WKT geometries
- Map column names to Darwin Core terms
- Check eventDate
- Dataset structure
- Data quality report

Examples on how to work with the package can be found here: https://obis.org/manual/processing/

Data harvest and DOIs

You will see more in detail how the data flows from EurOBIS, to EMODnet Biology, OBIS and GBIF as well as how to proceed to create a DOI for your dataset

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	Data harvest and DOIs
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:32 PM

Table of contents

Introduction

EMODnet Biology harvest procedures

Transfer from EurOBIS to the EMODnet Biology Portal

Data export to OBIS and GBIF

Updates and DOIs

Introduction

So you published your dataset on IPT and checked that there are no issues. Now it's time to contact bio@emodnet.eu and inform us that your dataset is ready to be harvested. Let's look again at the video from the introduction (In the "What is EMODnet Biology" book of this course) to see where your dataset will flow to:

EurOBIS data flow

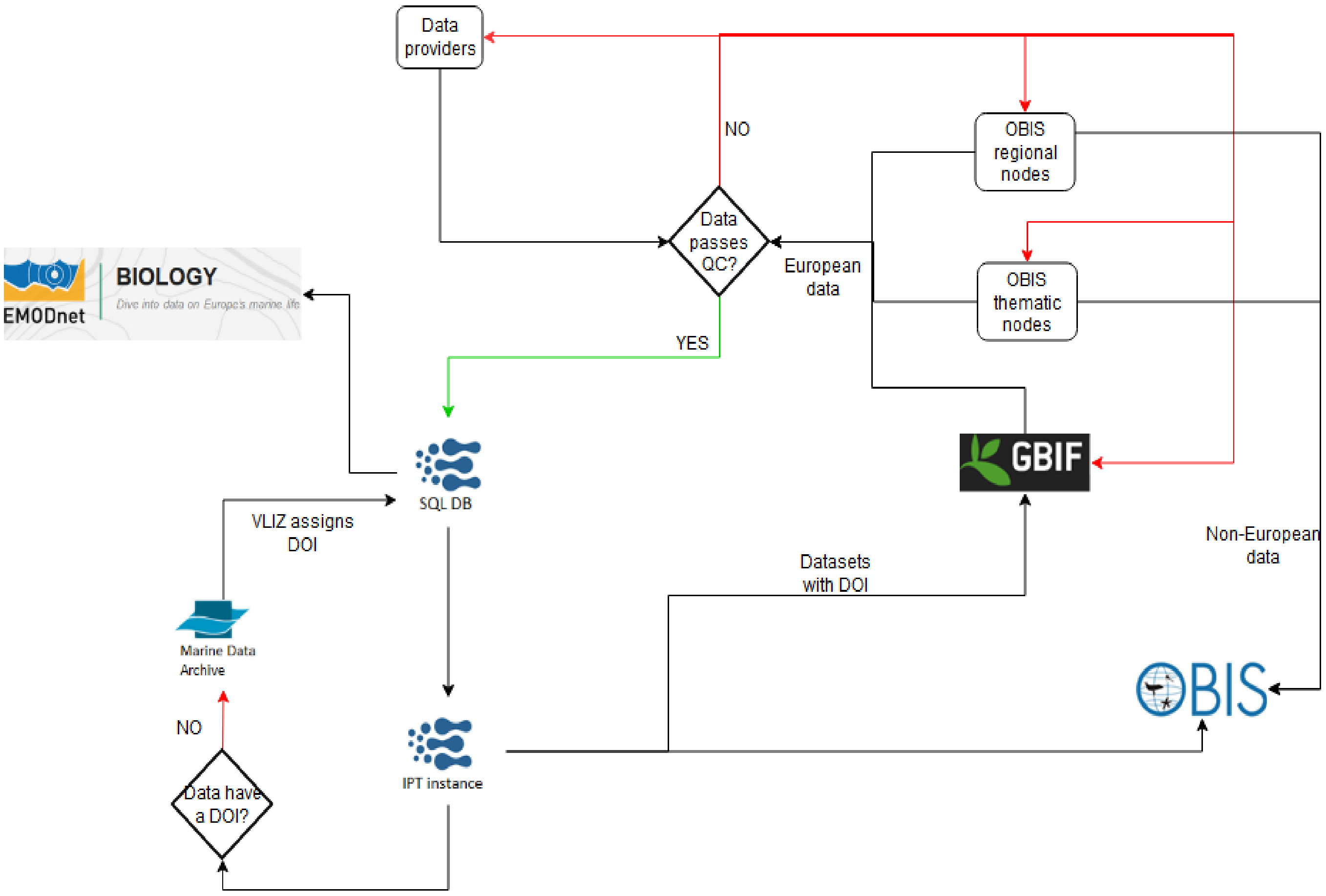
EMODnet Biology harvest procedures

After you send the email that the dataset is ready for harvest, the EMODnet Biology data management team will

1. Check if the metadata record in the EMODnet Biology catalogue meets requirements,
2. Perform a visual QC of the dataset (as explained in the "Technical metadata and data entity integrity" book of this course)
3. Run it again to the QC tool demonstrated in the "IPT datasets and Quality Control" book of this course.

If all is fine, the dataset will be listed as to be included in the next EMODnet Biology harvest.

Harvests by EMODnet Biology are a semi-automated process carried out on a periodical basis (about one harvest every 3 months). All datasets judged fit for harvesting will then be processed by automated procedures and become available through the EMODnet Biology Portal (both the viewer and the download toolbox). New IPT resources will be created on the EurOBIS IPT, from where your dataset can be accessed by OBIS and GBIF. Take a look at the more detailed data overview schema below. In the next pages, you will next go through all the different steps in detail.



Transfer from EurOBIS to the EMODnet Biology Portal

Datasets are harvested into the EurOBIS SQL database which serves as a staging database for export to the EMODnet Biology Postgres database. After a harvest cycle all EurOBIS data undergo automated standardisation procedures before being presented through the EMODnet portal. Integration of the EurOBIS data with Marine Regions and the World Register of Marine Species allows the EMODnet Biology download toolbox to include filters based on geographical areas (EEZ's, Territorials Seas, Marine Regions, MEOW's and IHO Sea Areas), taxonomic classifications and functional groups.

Additionally before being made available through the toolbox the harvested data will undergo a series of automated data transformations. These data transformations include:

- Flattening out event hierarchy
- Combining event and occurrence tables
- Standardisation of the measurement or facts data through the link with the BODC vocabularies
- Standardisation of all values of a measurement to a common unit
- Combining the standardised measurement or fact data and event/occurrence table
- Generating the entire taxonomic classification of the accepted taxon name based on the link with the World Register of Marine Species (through scientficNameID)

Below you can inspect (and download) the demo dataset as it will be available through the EMODnet download toolbox. The columns highlighted in yellow are generated.The columns highlighted in green are transformed by standardisation procedures involving the BODC vocabularies. Note that there are 2 tabs in the excel file: one containing the data and one containing the technical metadata for the additional measurement or fact parameters.

download data file

Data export to OBIS and GBIF

The EurOBIS IPT

EurOBIS is the European node of OBIS and as such the EurOBIS data management team is responsible for transforming datasets to be ready for upload to OBIS. EurOBIS maintains an IPT dedicated to this purpose; The "EurOBIS IPT" contains all datasets which EurOBIS sends to OBIS. As the EurOBIS database also includes data from other OBIS nodes (that also send data directly to OBIS), not all the datasets within the EurOBIS database (and EMODnet Biology) are included in the EurOBIS IPT. Once in the EurOBIS IPT, OBIS will harvest the datasets within hours and they will be directly available through OBIS at http://obis.org/.

The EurOBIS IPT also contains the datasets meant to be harvested by GBIF linked to the correspondent publisher ID. Before a dataset is sent to GBIF a DOI will be assigned to it by either the data provider or by the EurOBIS team.

The metadata used on the EurOBIS IPT is exported from the EMODnet Catalogue. There is no automated procedure in place to update the EMODnet Catalog records with the metadata you provide through your IPT. Therefore if you require changes to the dataset's metadata - apart from changing them in your IPT - you will also need to inform bio@emodnet.eu. The EurOBIS team can make the changes in the EMODnet Catalogue, which will then be updated automatically in all systems which are fed by the EurOBIS IPT.

The data used on the EurOBIS IPT is the data as stored in the EurOBIS SQL, which should be identical to the data presented to us through the IPT of the data provider.

Updates and DOIs

If all data and metadata are ok, EMODnet Biology will recommend to have a DOI for the dataset.

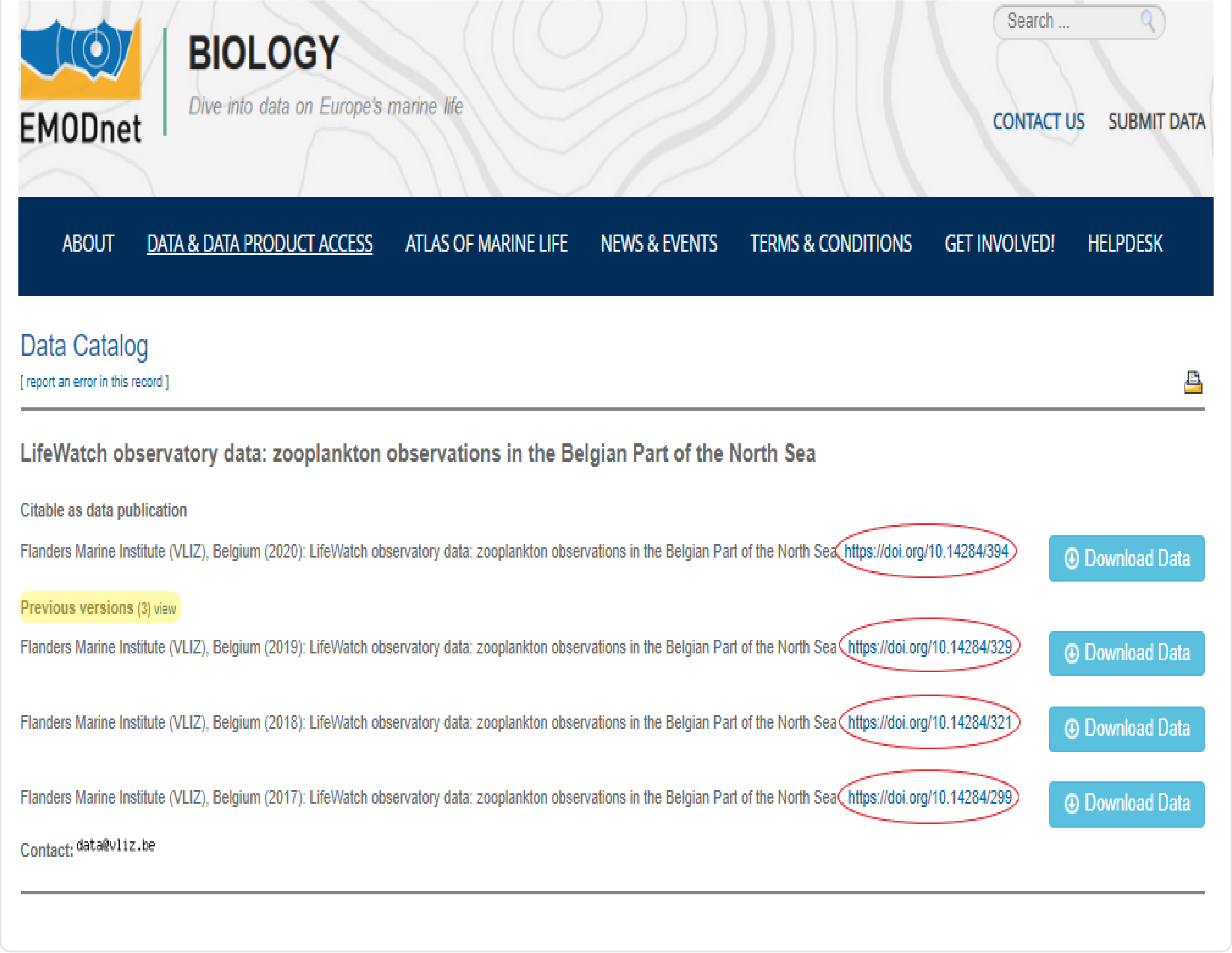
The DOI can be created either by the data provider or by the EMODnet Biology data management team:

- **By the data provider**: The DOI should be created after the QC procedures have been carried out by the EMODnet Biology team. In this case, the "Publisher of the dataset's citation" (already mentioned in the "Metadata in EMODnet Biology" book of this course) can be the data provider's organization or the institution/system that archives the DOI.

- **By the EMODnet Biology team**: We would contact the data providers after the EurOBIS data harvest with the proposal. If the DOI is assigned by EMODnet Biology, the "Publisher" will be the Marine Data Archive (MDA) as that is the publishing instance that will be responsible for the long term storage of that version of the dataset.

In the case of an updated dataset, after you create an new version of the resource you should inform bio@emodnet.eu. At this point we would run the QC procedure again and if there aren't new issues with the dataset, it would be re-harvested and a new DOI would be assign to this new version of the dataset.

If the DOI has been created by EMODnet Biology, all the versions of the dataset will be interlinked within IMIS, the metadata system behind the EMODnet Biology Catalogue (as explained in the "Metadata in EMODnet Biology" book of this course), as shown in the dataset *"LifeWatch observatory data: zooplankton observations in the Belgian Part of the North Sea"*:



Data Catalog

[report an error in this record]

LifeWatch observatory data: zooplankton observations in the Belgian Part of the North Sea

Citable as data publication

Flanders Marine Institute (VLIZ), Belgium (2020): LifeWatch observatory data: zooplankton observations in the Belgian Part of the North Sea https://doi.org/10.14284/394

Previous versions (3) view

Flanders Marine Institute (VLIZ), Belgium (2019): LifeWatch observatory data: zooplankton observations in the Belgian Part of the North Sea https://doi.org/10.14284/329

Flanders Marine Institute (VLIZ), Belgium (2018): LifeWatch observatory data: zooplankton observations in the Belgian Part of the North Sea https://doi.org/10.14284/321

Flanders Marine Institute (VLIZ), Belgium (2017): LifeWatch observatory data: zooplankton observations in the Belgian Part of the North Sea https://doi.org/10.14284/299

Contact: data@vliz.be

How to access data

Site:	OceanTeacher
Course:	Contributing datasets to EMODnet Biology
Book:	How to access data
Printed by:	Ruben Perez
Date:	Wednesday, 1 July 2020, 3:32 PM

Table of contents

1. Via EMODnet Biology
2. Via OBIS
3. Via GBIF

1. Via EMODnet Biology

There are several ways to find your data Via EMODnet Biology:

- Catalogue
- Download toolbox
- Web mapper
- IPT
- Webservices

The presentation briefly introduces the access points listed above.

2. Via OBIS

As EMODnet Biology data are shared with OBIS, it is also possible to search and download it using:

- Mapper
- Download toolbox
- rOBIS R package

The presentation briefly introduces the access points listed above.

3. Via GBIF

Due to the sharing of data via EurOBIS, EMODnet Biology data can also be found in GBIF:

- Catalogue

The presentation briefly introduces the access point listed above.