EMODnet

European Marine Observation and Data Network

EMODnet Thematic Lot n° VI - Biology

EMFF/2019/1.3.1.9/Lot 6/SI2.837974

Start date of the project: 19/04/2021 - (24 months)

EMODnet Phase IV

D2.6: Report on the standardisation and integration of the proposed new and updated datasets

The European Marine Observation and Data Network (EMODnet) is financed by the European Union under Regulation (EU) No 508/2014 of the European Parliament and of the Council of 15 May 2014 on the European Maritime and Fisheries Fund.

Disclaimer

Document info

Title	D2.6: Report on the standardisation and integration of the proposed new and updated datasets
WP title	WP2: Access to marine biological data
Task	T1: Maintain and improve a common method of access to data held in repositories
Authors [affiliation]	Leen Vandepitte, Ruben Perez Perez (VLIZ)
Dissemination level	Public
Submission date	30/05/2023
Deliverable due date	18/04/2023

Contents

D2.6: Report on the standardisation and integration of the proposed new and updated datasets

1 Introduction

During phase IV of the EMODnet Biology project (April 2021 – April 2023), the data management team at VLIZ has been in close contact with all partners involved in WP2, to ensure a continuous data flow to the EurOBIS database, the data backbone of EMODnet Biology.

As in previous phases, this phase has brought new data from the partners to EMODnet Biology, next to focusing on the expansion of currently available datasets, so they include related data such as e.g. abundance, biomass and abiotic measurements related to the species observations.

During the first six months of this EMODnet Biology 4 phase, the data management team had consulted with all WP2 partners, to have in-depth discussions about the data they had already delivered – with a focus on how these could be extended and improved – and to get insights in data that could additionally flow to EMODnet Biology (see: D2.2 Technical implementation of data flows for the new project partners/sub-contractor).

Phase IV of EMODnet Biology has delivered 174 new datasets from WP2 partners and additional updates of 48 pre-existing datasets, now available in the EMODnet Central portal viewer. Next to adding new data and data updates to the project, Work Package 2 was also set out to comply to several predefined standards, vocabularies and data formats within this project, to optimise the integration of scattered marine biological datasets. These are (1) the Darwin Core Archive standard, (2) the World Register of Marine Species (WoRMS) (standardised vocabulary) and (3) the BODC NVS2. Formatting to and complying with these standards and vocabularies allows interoperability not only within EMODnet Biology but also with other initiatives like OBIS. In this document, we are reporting on their implementation and general progress.

2 Darwin Core Archive

The Darwin Core Archive is an internationally recognised biodiversity informatics data standard that simplifies the publication of biodiversity data. During this phase of the EMODnet Biology project, we moved towards transitioning the existing and new data from Occurrence Core to Event Core, offering multiple advantages in terms of the richness and integrity of the measurements to be stored in combination with the biological occurrences as well as minimising data redundancy. Adhering to the Event Core standard at first sight seemed more complex compared to the Occurrence Core format to a number of partners, due to the fact that the data is now structured in 3 instead of 2 tables (Figure 1).

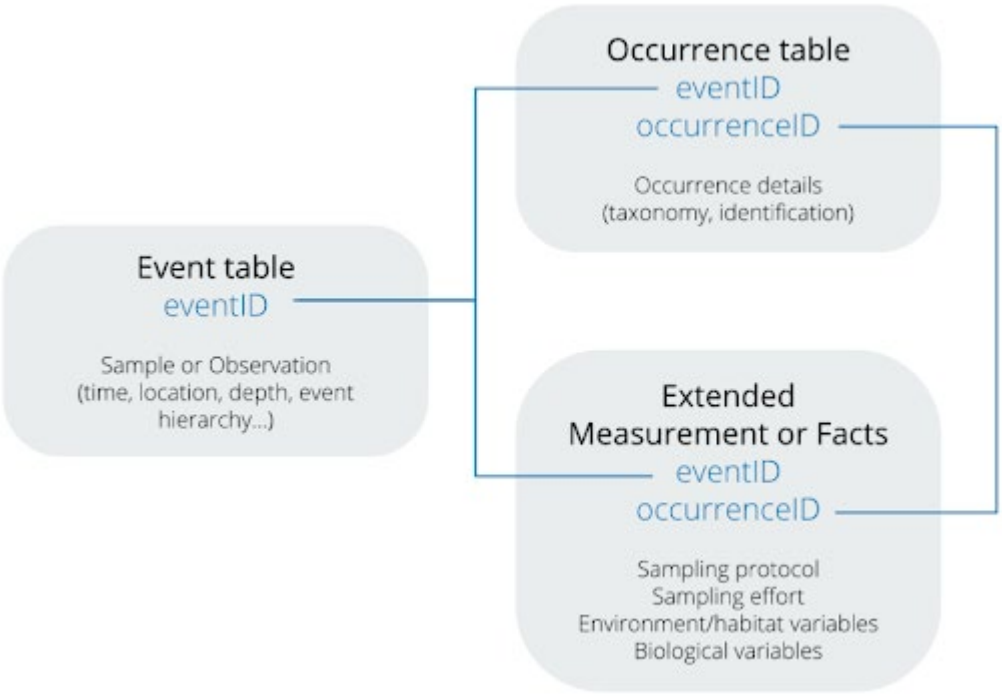

Figure 1. Darwin Core based schema used in EMODnet Biology

Intensive and continuous follow-up by the data management team however made sure that each partner understood the new format and how to apply it.

People within the partner institutes that were new to the field of data management, were all introduced to the online training course on data formatting, quality control and data publishing, which was developed during EMODnet Biology Phase 3.

In total, EMODnet Biology partners have delivered 222 datasets in the course of this phase (April 19th 2021 – April 18th 2023), of which 198 are available in the preferred Event Core format of the Darwin Core (DwC) standard. The other 24 datasets are currently available in Occurrence Core format and are either in the process of being transformed, under evaluation to transform or planned to be transformed. Transformation from Occurrence Core to Event Core format does not only imply reorganisation of the data from two tables to three tables, but it also implies de-duplication of data and information and the possibility to add specific data and information that could not be captured through the Occurrence Core format.

Although Event Core is the recommended and preferred format for data delivery and integration into the EurOBIS database, data occasionally provided in Occurrence Core format - provided from outside of the EMODnet Biology consortium – were and will still be accepted for validation and integration. Wherever possible, the data management team has assisted the original data provider in the transfer from Occurrence Core to Event Core, and will continue to do so in the future. During Phase IV, the data management team has already invested considerable time and effort in transferring existing datasets to Event Core, and in standardising the extended Measurements or Facts (eMoF) that were already present in the database, from datasets ingested in the pre-Event Core period. As the Occurrence Core had been the standard for more than 15 years in EurOBIS, a transfer to Event Core

for all available datasets cannot be done automatically overnight. Transferring to the Event Core format does not only require efforts on the data management side, but – wherever possible – also needs to involve the original providers, as they can also benefit from this transfer, and extra data – which were not captured through Occurrence Core – could now also be added.

3 World Register of Marine Species (WoRMS)

The World Register of Marine Species (WoRMS) is the authoritative and comprehensive list of names of marine organisms worldwide and is the taxonomic backbone of EMODnet Biology, next to being a central part of the LifeWatch Species Information Backbone.

Within EMODnet Biology, the data management team strives to match as many taxon names to WoRMS as possible, in close collaboration with the EMODnet Biology data providers. This does not just allow quality checks on the used taxon names (e.g. spelling of taxon names), but it also greatly improves the interoperability of the data.

At the end of Phase IV, the EurOBIS database held 103,613 species names linked to the World Register of Marine Species (WoRMS), of which 73,029 are accepted species names. Of these accepted species names, 27,959 were documented in the European Register of Marine Species (ERMS). Two factors can explain this large difference in total number of marine species versus documented European marine species:

1. European data providers do not limit their sampling campaigns, research and monitoring to European marine waters. As they share their full datasets, occurrence information from outside of Europe also finds its way to the EurOBIS database and subsequently, EMODnet Biology. It is better to keep a dataset as a whole – even if it contains some data from outside of the European marine waters – than to split it up, with the risk of both parts becoming permanently disconnected.
2. Just as the World Register of Marine Species, the European Register of Marine Species is a dynamic work-in-progress. A number of species can already be documented in WoRMS, but with incomplete distribution information, thus not (yet) being displayed as part of the European marine waters. WoRMS relies on voluntary contributions of taxonomic and thematic editors– and all its sub-registers -, which can cause some (temporary) gaps in the available information.

	April 2021	**April 2022**	**April 2023**
Datasets	1,077	1,197	1,233
Species names	98,024	103,465	103,613
Accepted species names	75,658	71,620	73,029

With each data harvest cycle, occurrence information is being added to EurOBIS for species that are documented for the first time within the database.

Since the start of Phase 4 (April 2021), no less than 1,708 accepted species names were reported for the first time in the database with one or more observations. This indicates that – although we know about the existence of a species, as its name is documented in WoRMS/ERMS – EurOBIS is capturing an actual field observation of these species for the very first time. These first records of a species in EurOBIS are extremely important, as they can help to gain new insights in the distribution of certain species, and also help to more clearly identify remaining gaps: if a species has been observed once, it is most likely that other observations of this species are being made, but the datasets in which they are contained have not yet found their way to EurOBIS/EMODnet Biology.

Approximately 22,212 taxon names are currently not yet matched to the World Register of Marine Species (WoRMS). The matching process is continuously ongoing. To solve non-matching names, the data management team is often dependent on the WoRMS taxonomic experts, who contribute to WoRMS on a voluntary basis. A number of these names represent terrestrial and/or freshwater taxa, which were 'accidentally' caught in e.g. coast-related datasets and were still kept in the datasets, as matter of completeness.

During July 2021, we experienced a decrease on the number of accepted species names in EurOBIS of 4,890 names. Although the cause of this decrease is still being investigated and further updates will be reported in the Phase V quarterly reports, there are several reasons that could explain it; for example, the dynamic quality control nature of WoRMS and EurOBIS which constantly clean records and optimise themselves, potentially causing a decrease of records to account for de-duplication.

4 BODC controlled vocabularies

The BODC controlled vocabularies are a collection and lists of standardised terms of relevance for ocean sciences, which are being used to label data to enable interoperability and overcome ambiguities.

Within the EurOBIS database, the eMoF table is the place where data and information beyond the 'what-where-when' of a species observation is captured, such as e.g. the abundance, biomass and length of the species, as well as accompanying environmental measurements and details on the used sampling gears and protocols, either linked to a specific occurrence, or to the sampling event in general. Data and information in this extended Measurements or Facts (eMoF) extension table is mostly being mapped to the BODC Parameter Usage Vocabulary (PUV) or P01 collection. The P01 collection is a controlled vocabulary for labelling measured and observed variables in data files and databases. Linking the data and information in the eMoF table to P01 ensures the highest level of standardisation within the database. Next to the P01 collection, other BODC vocabulary collections are being used to standardise terms related not only to the sampling facts and measurements (P01) and their units (P06), but also to biotic facts such as the S11- Lifestage and S10- Gender collections, to habitat related facts such as their M21 collection and to sampling related information such as the L22, L05, C17 and Q01 collections.

Mapping to the BODC collections was introduced with the start of the previous Phase (Phase 3 - April 2019). Back then, this introduction was accompanied by a learning curve, both on the data management and the data provider side. The data management team is still continuously assisting all data providers in how to adhere to this standard, how to select the appropriate terminology from the BODC collections, and how to implement this in their data flows.

5 In summary

EMODnet Biology has grown tremendously over the last 2 years, both in terms of making available new datasets as well as ensuring that the data they contain adhere to international standards. This standardisation has greatly improved the FAIR-level (Findability-Accessibility-Interoperability-Reproducibility) of the available data.

Phase IV has also seen the first introduction of image-based datasets. In addition, the first steps were made behind the scenes to be able to include DNA-derived data into the system.

The full overview of available datasets in EMODnet Biology can currently be found at https://www.eurobis.org/dataset_list