EMODnet

European Marine Observation and Data Network

EMODnet Biology

EASME/EMFF/2017/1.3.1.2/02/SI2.789013

Start date of the project: 19/04/2019 - (24 months)

EMODnet Phase III

D2.6: Report on data standardization of proposed new and update 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 data standardization of proposed new and update datasets
WP title	WP2: Data access to marine biological data
Task	T1: A common method of access to data held in repositories
Authors [affiliation]	Leen Vandepitte, Ruben Perez Perez (VLIZ)
Dissemination level	Public
Submission date	15/04/2021
Deliverable due date	18/04/2021

Contents

D2.4: Report on data standardization of proposed new and update datasets

1 Introduction

During the renewal phase of the EMODnet Biology 3 project (April 2019 – April 2021), 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, as backbone of the EMODnet Biology data portal.

This renewal 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. In addition, new datasets from data grant partners also flowed to EMODnet Biology.

During the first six months of this EMODnet Biology 3 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.

The renewal phase of EMODnet Biology 3 has delivered 223 new datasets from WP2 and WP3 partners and data grant partners, and additional updates of 63 pre-existing datasets available in the portal. Next to adding new data and data updates to the project, Work Package 2 was also set out to comply to a number of predefined standards, vocabularies and data formats within this project, to optimize the integration of scattered marine biological datasets. These are (1) the Darwin Core Archive standard, (2) the World Register of Marine Species (WoRMS) (standardized vocabulary) and (3) the BODC controlled vocabularies. Formatting to and complying with these standards and vocabularies allows interoperability with the EMODnet Biological Portal. In his document, we are reporting on their implementation, including general progress and difficulties encountered.

2 Darwin Core Archive

The Darwin Core Archive is an internationally recognized biodiversity informatics data standard that simplifies the publication of biodiversity data. In particular, 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 occurrences. 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. 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.

An online organised, hands-on workshop on data formatting, quality control and data publishing (M12) has enabled the data management team to thoroughly explain the Event Core format to all data partners – pre-existing and new – and the data grant holders of this phase, making the data processing and delivery process as smooth as possible. This hands-on workshop was originally planned to take place in person, at the Flanders Marine Institute (VLIZ), but had to be transformed to an online workshop due to the global COVID-

19 pandemic. During the whole renewal period, the data management team has kept close contact with all data providers, ensuring a good follow-up of data processing, formatting, quality control and delivery. As this could not be done in person, the data management team has made use of all online available tools (email and video-conferences), to guide providers through the formatting process, and helping them understand all the details linked to the DarwinCore Event Core format.

In total, WP2 partners, WP3 partners and data grant holders have delivered 223 new datasets in the course of this phase (April 19th 2019 – April 18th 2021), of which 204 are available in the preferred Event Core format of the Darwin Core (DwC) standard. For the 19 datasets still in Occurrence Core, 14 were retrieved from museum collections in the pre-existing Occurrence Core format, where there was no immediate benefit in transferring them from Occurrence Core to Event Core, as this would by no means give rise to additional or better data and information. The other 5 datasets in Occurrence Core format 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 reorganization of the data from two tables to three tables, 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 the renewal phase, the data management team has already invested considerable time and effort in transferring existing datasets to Event Core, and in standardizing the extended Measurements or Facts 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 about 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 the EMODnet Biology Data Portal, 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 III, the EurOBIS database held 98,024 species names linked to the World Register of Marine Species (WoRMS), of which 75,658 accepted species names. Of these accepted species names, 28,068 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 EurOBIS. 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.

	May 2017	May 2018	May 2019	May 2020	April 2021
Datasets	746	875	899	1,026	1,077
Species names	72,727	77,733	80,202	94,631	98,024
Accepted species names	58,729	60,868	62,220	73,706	75,658

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 the renewal phase of EMODnet Biology 3 (April 2019), no less than 7,155 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.

Approximately 21.525 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.

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 in order 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) 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 L22, L05 and S10 collections, to

habitat related facts such as their M21 collection and to sampling related information such as the C17 and Q01 collections.

As mapping to the BODC collections has only been introduced with the start of this renewal phase (April 2019), this introduction is accompanied by a learning curve, both on the data management and the data provider side. The data management team assisted 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.

At the time of reporting, 23,818,715 records in the eMoF table were linked to 565 unique BODC terms from the various BODC collections.

5 In summary

Tremendous progress was made over the last 2 years in terms of making available new datasets, and ensuring that all the data they contain are adhering to international standards. This standardization greatly improves the FAIR-level (Findability-Accessibility-Interoperability-Reproducibility) of the available data.

This renewal phase of the EMODnet Biology 3 project (April 2019 – April 2021) has been characterized by some far-reaching circumstances. In first instance, all partners needed to shift from providing data in Occurrence Core to Event Core, which was not always straightforward but overall has been implemented quite smoothly. Next to the format-adaptation, partners were also introduced to the BODC controlled vocabularies. This involved learning about the concept of these vocabularies, how to search them and how to make sure specific terminology from these vocabularies is unambiguously linked to the information in the datasets. As has happened in the past, occasional changes in personnel with the partners has given some temporary delays as new staff members had to get familiar with the data and the way they needed to be formatted for submission to EMODnet Biology. Lastly, the past year gave rise to a situation never seen before, the global COVID-19 outbreak, confining people to working from home and adapting to never-seen circumstances. This has proven to be the biggest challenge of all for the data providing partners, causing temporary delays in the overall progress of data processing, formatting and delivery.

Despite the mentioned challenges and difficulties, the renewal phase of the EMODnet Biology 3 project has taken tremendous leaps forward, both on the level of data availability and adhering those data to internationally accepted data standards, data formats and vocabularies. All this gives enormous trust to move forward towards the next phase, and to keep up the quality and FAIR-ness of the data within the EMODnet Biology project.