EMODnet

European Marine Observation and Data Network

EMODnet Thematic Lot n° V - Biology

CINEA/EMFAF/2022/3.5.2/SI2.895681

Start date of the project: 10/05/2023 - (24 months)

EMODnet Phase V

Operational Phase

D2.1.2: Report on the standardisation and integration of datasets published during the Phase

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	D2.1.2: Report on the standardisation and integration of datasets published during the Phase
WP title	WP2: Data management
Task	T1: Maintain and improve a common method of access to data held in repositories
Authors [affiliation]	Ruben Perez Perez, Leen Vandepitte (VLIZ)
Dissemination level	Public
Submission date	07/05/2025
Deliverable due date	09/05/2025

Contents

Report on the standardisation and integration of datasets published during the Phase

1 Introduction

During phase V of EMODnet Biology (May 10th 2023 – May 9th 2025), the data management team at VLIZ has been in close contact with all consortium organisations involved in WP2 to ensure a continuous data flow to the data infrastructure backbone of EMODnet Biology. The main objective of WP2 is defined by EMODnet Biology Task 1 - "Maintain and improve a common method of access to data held in repositories"

As in previous phases, this phase has brought new data from the consortium organisations 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 V phase, the data management team had consulted with all WP2 organisations, 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: D2.2.1. Summary on consortium data flows).

Phase V of EMODnet Biology has delivered a total of 303 datasets, of which 222 are new datasets and 81 are updates of pre-existing datasets, now available in the EMODnet 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 British Oceanographic Data Centre and Natural Environment Research Council Vocabulary Server (BODC NVS2). Formatting to and complying with these standards and vocabularies allows interoperability not only within EMODnet Biology but also with other initiatives like the Ocean Biodiversity Information System (OBIS). In this document, we are reporting on their implementation and general progress.

2 Darwin Core Archive

The Darwin Core Archives are an internationally recognised biodiversity informatics data standard that simplifies the publication of biodiversity data. During this phase of EMODnet Biology, we continued moving towards transitioning the existing and new data from Occurrence Core to Event Core when appropriate, 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 organisations, due to the fact that the data is now structured in 3 instead of 2 tables (Figure 1).

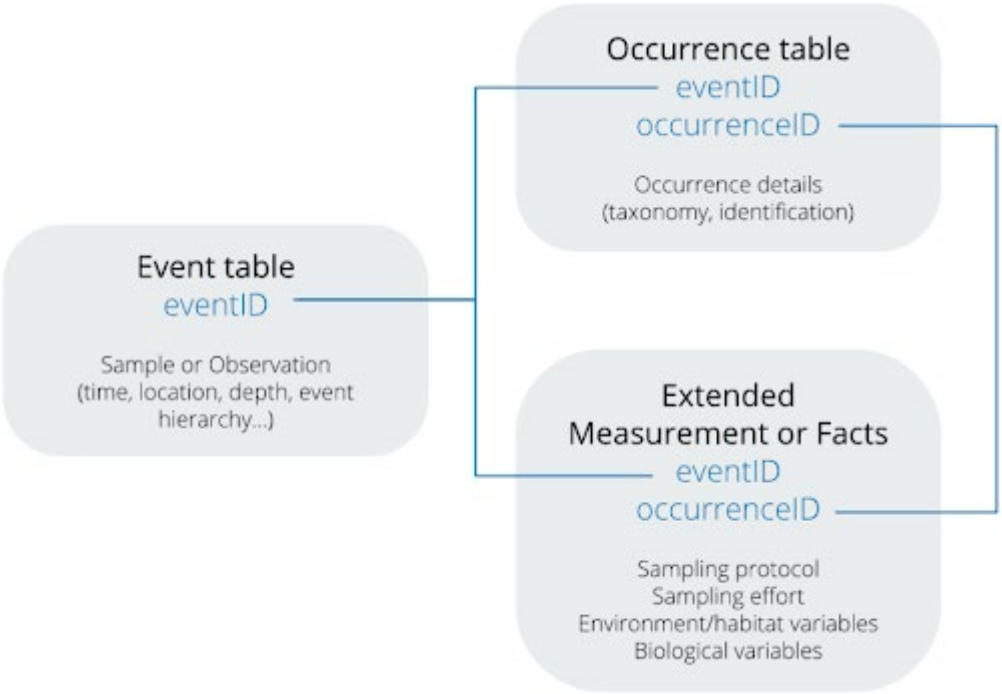

Figure 1. Darwin Core based schema used in EMODnet Biology

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

People within the 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 and revamped during EMODnet Biology phase IV.

In total, EMODnet Biology WP2 organisations have delivered 303 datasets in the course of this phase (May 10th 2023 – May 09th 2025), of which 150 are available in Event Core format of the Darwin Core (DwC) standard. The other 153 datasets are currently available in Occurrence Core format which is adequate for datasets with simpler sampling methodologies. 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 continue to 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.

In addition, Darwin Core has also grown to accommodate environmental DNA (eDNA) derived occurrences data in an interoperable manner. In this regard the "DNA derived data Extension" table

was incorporated to the standard and integrated into the database, which included several modifications and new features needed for the database and data portal. A deliverable capturing the main guidance and information regarding this type of data is available via the link. By the end of Phase V, a total of 3 dataset have successfully been integrated into EMODnet Biology.

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 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 V, the database held 94,835 species names linked to the World Register of Marine Species (WoRMS), of which 75,387 are 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 the 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 (like ERMS)-, which can cause some (temporary) gaps in the available information.

	April 2021	**April 2023**	**April 2025**
Datasets	1,077	1,233	1,450
Species names	98,024	103,613	94,835
Accepted species names	75,658	73,029	75,387

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

Since the start of Phase V (May 10th 2023), no less than 1,873 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 – the data in the database is capturing an actual field observation of these species for the very first time. These first records of a species in the database that hosts the data published in EMODnet Biology, 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 EMODnet Biology.

Approximately 21,921 taxon names are currently not yet matched to the World Register of Marine Species (WoRMS). The matching process is continuously ongoing and is done by the WoRMS team. To solve non-matching names, the EMODnet Biology 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.

Note that the numbers in the table and paragraphs above do not include the last data harvest of the Phase V period as the publication process had not been completed by the time this deliverable was published.

4 BODC NVS2 controlled vocabularies

The British Oceanographic Data Centre and Natural Environment Research Council Vocabulary Server (BODC NVS2) 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 database that host data published in EMODnet Biology, the extended Measurements or Facts (eMoF) table is the place where data and information beyond the 'what-where-when' of a species observation is captured. This table includes information such as e.g. organism quantifications (abundance, biomass, concentration, etc.), biometrics (length, width, etc.) organism characteristics (life stage, sex, size, behaviour, etc.), 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, which 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 and C17 collections.

Mapping to the BODC collections was introduced with the start of Phase III (April 2019 and the data management team is continuously assisting all data providers in how to adhere to this standardised approach, how to select the appropriate terminology from the different 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 FAIRness (Findability-Accessibility-Interoperability-Reproducibility) of the available data.

Phase V has also seen the first successful introduction of DNA-derived data into the system and a consolidated increase of recently incorporated data sources such as occurrences derived from imaging, tracking and acoustic telemetry.

The full overview of available datasets in EMODnet Biology can currently be found in the EMODnet catalogue.