
# EMODnet Thematic Lot n° V- Biology // D2.5.1 Guidance for data management practices applied to omics data

**CINEA/EMFAF/2022/3.5.2/SI2.895681**

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

**Operational Phase**

## Guidance for data management practices applied to omics data [D2.5.1]

*The European Marine Observation and Data Network (EMODnet) is financed by the European Union under Regulation (EU) 2021/1139 of the European Parliament and of the Council of 7 July 2021 establishing the European Maritime, Fisheries and Aquaculture Fund.*

## Disclaimer

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the CINEA or of the European Commission. Neither the CINEA, nor the European Commission, guarantee the accuracy of the data included in this study. Neither the CINEA, the European Commission nor any person acting on the CINEA's or on the European Commission's behalf may be held responsible for the use which may be made of the information.

## Document info

| | |
|---|---|
| Title (and reference) | Deliverable Guidance for data management practices applied to omics data [D2.5.1] |
| WP title (and reference number) | WP2- Data management |
| Task (and reference number) | Task 1: Maintain and improve a common method of access to data held in repositories |
| Authors [affiliation] | Lynn Delgat (VLIZ) |
| Dissemination level | Public |
| Submission date | 2025-09-07 |
| Deliverable due date | 2025-09-09 |

# Contents

# D2.5.1 Guidance for data management practices applied to omics data

Biodiversity data, once obtained only through direct visual observation of the different taxa, are now obtained through different methods or techniques. The biodiversity data community continuously enhances and expands the data standards in order to allow for appropriate description and information capture that will allow the wider community to reuse the collected data. In EMODnet Biology the same process applies, the data infrastructure that hosts the published data was initially designed to incorporate (visual) taxon observations, but has since been expanded to accommodate for data from imaging and tracking methods.

The developments described in this document took place during EMODnet Biology's Phase V (2023-09-10 to 202-05-09) and allowed for the management and publication of taxon occurrences derived from molecular methods. These developments were not only needed to align with other biodiversity repositories but were also a requirement from the funding organisation "The service shall evolve to include omics data for the marine environment". The outcome during the last two years was the publication of five datasets, with many more expected in the near future as providers become aware of this work and are willing to make their data available.

## 1 DNA derived data

Biodiversity data is being generated by molecular methods at an increasing rate, generating large amounts of data. Most molecular methods result in the generation of sequence data. These sequences are stored in sequence databases such as NCBI's Sequence Read Archive (SRA) and European Nucleotide Archive (ENA). By being stored only in sequence databases, in many cases as compressed files containing raw data that requires extensive bioinformatic analysis, this data is mainly re-used by other researchers involved in molecular research; however, the information that can be derived from these sequences can be extremely useful to the wider scientific community. To unlock this data in a broader context, biodiversity databases such as GBIF and OBIS are encouraging molecular researchers to submit the occurrences derived from DNA data to biodiversity databases. The occurrences derived from DNA data serve as a valuable complement to the biodiversity data collected by traditional methods, not only due to the large amount of data being generated by molecular methods, but also because it allows recording data on currently undescribed species, and the detection of taxa that are difficult or even impossible to detect through traditional methods.

To accommodate the publication of occurrence data derived from DNA, a new Darwin Core extension was added to the Darwin Core standard: the DNA derived data extension. This extension contains terms from the Minimum Information about any (X) Sequence (MIxS), Global Genome Biodiversity Network (GGBN), and Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) standards. To facilitate the formatting of DNA derived data using the Darwin Core standard, a guide was created by GBIF: Publishing DNA-derived data through biodiversity data platforms. In an update of this guide, a paragraph was added by OBIS to highlight some specifics related to marine DNA derived data, such as using LSIDs from the World Register of Marine Species (WoRMS) for the scientificNameID and adding the taxon identifiers from the reference databases (e.g. Barcode index numbers: BINs from BOLD) in the taxonConceptID field.

## 2 Developments to allow integration of DNA derived data into EMODnet Biology

Given that DNA derived data is formatted using the new DNA derived data extension, which contains many new terms, the database hosting the data published in EMODnet Biology required to be updated to allow for the integration of this information. Given the large amount of terms available in the extension (i.e. 119) the

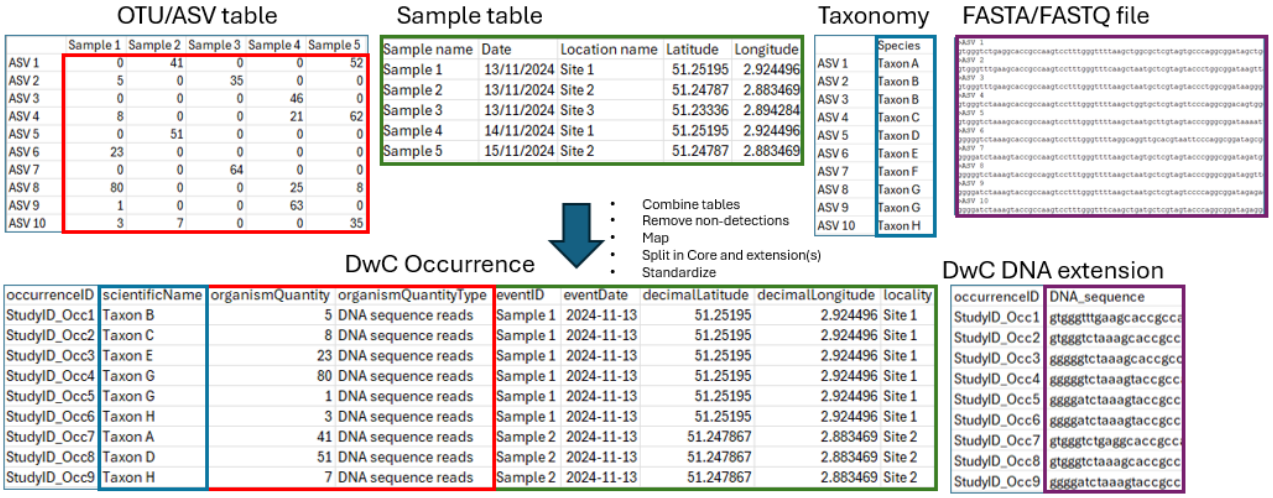
data management team in EMODnet Biology opted to use only a subset of those terms as fields in the database, for performance reasons. To decide which terms should be captured as fields in the database, a survey was conducted in December 2023 to be informed of the user perspective on which terms would be used or queried most frequently. In this survey, respondents (i.e. partners from EMODnet Biology or DTO-BioFlow task 3.1) could rate which fields they deemed important to document in a biodiversity database (from not needed to required) and which of those fields they would use to query, analyse or understand the data and whether that would be on an occasional or frequent basis. Based on the results of this survey, a set of nine fields was selected to be captured in a new "DNA" table in the database. The current priority was to accommodate enriched occurrences, and occurrences derived from (meta)barcoding data, so this set can later be expanded to accommodate other types of DNA derived data such as (meta)genomics. All other terms from the TDWG DNA extension are captured in a second DNA table, which is structured in a similar way to the eMoF table, i.e. a long format. In addition, the EMODnet Biology data management team (DMT) have adapted the database and procedures to enhance DNA derived data exposed through EMODnet Biology in two ways: (1) by providing a field "DNA derived" which can be TRUE or FALSE, to distinguish between enriched occurrences (i.e. DNA is additional evidence, on top of a specimen or observation) and DNA derived occurrences (i.e. DNA is the only evidence of the occurrence), and (2) by providing a standardized gene name, to allow querying on gene (since the source target_gene is free text).

In addition, work was done on improving the standards related to DNA derived data. For example, initially the DNA extension could only be used with an occurrence core if one wanted to capture sequences. The core table is the 'central' table of a Darwin Core Archive, only this core table has links to other tables in the archive while the other tables ('extensions') can only link to the core table. However, with an occurrence table as the core, the eMOF table cannot be used in the optimal way since it cannot be linked to an event table and thus all measurement or facts would need to be recorded at the occurrence level rather than the event level (i.e. resulting in a lot of data duplication). Therefore, the addition of an "occurrenceID" field to the DNA extension was requested to GBIF, to allow using the DNA extension with an Event Core while still allowing to link DNA sequences to occurrences (i.e. circumventing the DwC Star Schema in a similar way as was done in the eMOF extension). This new field was added in July 2024. In addition, the DMT have worked on updating the definitions / requesting new terms in the MIxS standard through discussions through GitHub to optimize their use in the context of metabarcoding, a work that is still in progress.

# 3 Guidelines for formatting and publishing of DNA derived data

### 3.1.1 Formatting DNA derived data: general

The formatting of DNA derived data follows the same concepts of formatting data from other sources. The main difference is that an additional extension is needed: the DNA derived data extension. In addition, the initial data files look different than those resulting from other methods. The initial data files will typically consist of several files: an OTU/ASV table, a sample table, a taxonomy table and a FASTA/FASTQ file. The first step of the data transformation is merging these tables into one file based on the common IDs (i.e. ASV/OTU and sample identifiers). Non-detections (read count = 0) can be removed after the merging. From that point, the general data transformation steps apply: mapping, standardizing of values, enhancing, splitting the information in core and extension tables, etc. A visual representation of the process (without enhancement) can be found in Figure 1. To enhance the dataset, additional fields are added, especially considering highly recommended/required fields (see section 4.1.4), providing more information on the dataset/methodology. The following paragraphs provide more detailed information on how you can perform the data transformation.




| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 |
|---|---|---|---|---|---|
| ASV 1 | 0 | 41 | 0 | 0 | 52 |
| ASV 2 | 5 | 0 | 35 | 0 | 0 |
| ASV 3 | 0 | 0 | 0 | 46 | 0 |
| ASV 4 | 8 | 0 | 0 | 21 | 62 |
| ASV 5 | 0 | 51 | 0 | 0 | 0 |
| ASV 6 | 23 | 0 | 0 | 0 | 0 |
| ASV 7 | 0 | 0 | 64 | 0 | 0 |
| ASV 8 | 80 | 0 | 0 | 25 | 8 |
| ASV 9 | 1 | 0 | 0 | 63 | 0 |
| ASV 10 | 3 | 7 | 0 | 0 | 35 |

| Sample name | Date | Location name | Latitude | Longitude |
|---|---|---|---|---|
| Sample 1 | 13/11/2024 | Site 1 | 51.25195 | 2.924496 |
| Sample 2 | 13/11/2024 | Site 2 | 51.24787 | 2.883469 |
| Sample 3 | 13/11/2024 | Site 3 | 51.23336 | 2.894284 |
| Sample 4 | 14/11/2024 | Site 1 | 51.25195 | 2.924496 |
| Sample 5 | 15/11/2024 | Site 2 | 51.24787 | 2.883469 |

| | Species |
|---|---|
| ASV 1 | Taxon A |
| ASV 2 | Taxon B |
| ASV 3 | Taxon B |
| ASV 4 | Taxon C |
| ASV 5 | Taxon D |
| ASV 6 | Taxon E |
| ASV 7 | Taxon F |
| ASV 8 | Taxon G |
| ASV 9 | Taxon G |
| ASV 10 | Taxon H |

| occurrenceID | scientificName | organismQuantity | organismQuantityType | eventID | eventDate | decimalLatitude | decimalLongitude | locality |
|---|---|---|---|---|---|---|---|---|
| StudyID_Occ1 | Taxon B | 5 | DNA sequence reads | Sample 1 | 2024-11-13 | 51.25195 | 2.924496 | Site 1 |
| StudyID_Occ2 | Taxon C | 8 | DNA sequence reads | Sample 1 | 2024-11-13 | 51.25195 | 2.924496 | Site 1 |
| StudyID_Occ3 | Taxon E | 23 | DNA sequence reads | Sample 1 | 2024-11-13 | 51.25195 | 2.924496 | Site 1 |
| StudyID_Occ4 | Taxon G | 80 | DNA sequence reads | Sample 1 | 2024-11-13 | 51.25195 | 2.924496 | Site 1 |
| StudyID_Occ5 | Taxon G | 1 | DNA sequence reads | Sample 1 | 2024-11-13 | 51.25195 | 2.924496 | Site 1 |
| StudyID_Occ6 | Taxon H | 3 | DNA sequence reads | Sample 1 | 2024-11-13 | 51.25195 | 2.924496 | Site 1 |
| StudyID_Occ7 | Taxon A | 41 | DNA sequence reads | Sample 2 | 2024-11-13 | 51.247867 | 2.883469 | Site 2 |
| StudyID_Occ8 | Taxon D | 51 | DNA sequence reads | Sample 2 | 2024-11-13 | 51.247867 | 2.883469 | Site 2 |
| StudyID_Occ9 | Taxon H | 7 | DNA sequence reads | Sample 2 | 2024-11-13 | 51.247867 | 2.883469 | Site 2 |

| occurrenceID | DNA_sequence |
|---|---|
| StudyID_Occ1 | gtgggtttgaagcaccgcca |
| StudyID_Occ2 | gtggggtctaaagcaccgcc |
| StudyID_Occ3 | gggggtctaaagcaccgcc |
| StudyID_Occ4 | gggggtctaaagtaccgcc |
| StudyID_Occ5 | ggggatctaaagtaccgcc |
| StudyID_Occ6 | ggggatctaaagtaccgcc |
| StudyID_Occ7 | gtggggtctgaggcaccgcca |
| StudyID_Occ8 | gtggggtctaaagcaccgcc |
| StudyID_Occ9 | ggggatctaaagtaccgcc |

*Figure 1. Illustration of part of the DwC data transformation of metabarcoding data.*

### 3.1.2 Formatting using the metabarcoding toolbox

For metabarcoding data, the data provider can use the user friendly Metabarcoding Data Toolkit (MDT) to help perform the data transformation to Darwin Core. All details on how to use the MDT are available in the MDT user guide. Using the MDT to perform the data transformation is useful in cases where data providers have a singular dataset to transform, and prefer performing the transformation through a graphical user interface rather than performing the transformation directly.

When using the MDT, it is possible to publish to GBIF through the MDT, alternatively, the provider can use the MDT to obtain the Darwin Core Archive file and then publish through other methods such as the IPT.

There are different MDT installations available (overview available here). Some allow publishing and some are for conversion only (e.g. GBIF Conversion MDT). Similar to IPT instances, data providers should choose the most relevant MDT based on their affiliation, region, project, or dataset.

### 3.1.3 Formatting without the metabarcoding toolbox

For other methods than metabarcoding, there is no tool available. Also, when dealing with multiple datasets resulting from the same bioinformatics pipeline (i.e. initial data files in the same format), it is beneficial to perform the data transformation oneself through coding, as one could then easily repeat the data transformation process for each dataset. The data transformation and publication follow the same procedures as data collected through other methods, there are however some specific things to consider when working with DNA data. There will be additional/other required/recommend fields, the additional DNA extension, and specific information needs to be mapped to specific DwC terms. All needed information on how DNA derived data should be formatted is available in the "Publishing DNA-derived data through biodiversity data platforms" guide. Additional interesting resources are the OBIS Manual: DNA derived data and the chapter on DNA derived data in the Ocean Teacher Biological Data Management Course.

### 3.1.4 EMODnet Biology recommendations

The table below shows the fields recommended by EMODnet Biology for capturing information of enriched occurrences and (meta)barcoding data (on top of the fields already recommended in the general EMODnet Biology guidelines).

*Table 1. EMODnet Biology recommendations for DNA related fields for enriched occurrences and (meta)barcoding data*

| Field name | Extension | Recommendation for enriched occurrences | Recommendation for (meta)barcoding data |
|---|---|---|---|
| **associatedSequences** | occurrence | highly recommended | required |
| **organismQuantity** | occurrence | | highly recommended |
| **organismQuantityType** | occurrence | | highly recommended |
| **sampleSizeValue** | occurrence | | highly recommended |
| **sampleSizeUnit** | occurrence | | highly recommended |
| **identificationRemarks** | occurrence | | highly recommended |
| **identificationReferences** | occurrence | | highly recommended |
| **taxonID** | occurrence | highly recommended if DNA_sequence is not provided | highly recommended if DNA_sequence is not provided |
| **taxonConceptID** | occurrence | | highly recommended |
| **verbatimIdentification** | occurrence | | recommended |
| **materialSampleID** | occurrence | | highly recommended |
| **DNA_sequence** | DNA | highly recommended | highly recommended |
| **target_gene** | DNA | required | required |
| **target_subfragment** | DNA | highly recommended if applicable | highly recommended if applicable |
| **pcr_primer_forward** | DNA | highly recommended | highly recommended |
| **pcr_primer_reverse** | DNA | highly recommended | highly recommended |
| **pcr_primer_name_forward** | DNA | highly recommended | highly recommended |
| **pcr_primer_name_reverse** | DNA | highly recommended | highly recommended |
| **pcr_primer_reference** | DNA | highly recommended | highly recommended |
| **seq_meth** | DNA | highly recommended | highly recommended |
| **otu_class_appr** | DNA | | highly recommended |
| **otu_db** | DNA | | highly recommended, if applicable (i.e. when (closed/open-) reference OTU clustering was used) |
| **sop** | DNA | | recommended |

# 4 DNA derived datasets in EMODnet Biology

Three metabarcoding datasets were included the November 2024 data publication, originating from the ARMS-MBON monitoring network, which deploys Autonomous Reef Monitoring Structures (ARMS) and

makes genetic assessments of the lifeforms that settled on these ARMs units during their deployment. The three datasets represent three analyses of different genetic markers: ITS, COI and 18S.
One dataset containing enriched occurrences was included in the April 2025 data publication, and will be available shortly after the publication of this report. This dataset originated from the MONICEPH project, which stores and analyses samples of cephalopods collected at the sea surface of the Azores, mostly during touristic activities. Sequencing of the COI marker was done as an additional source of information, providing more confidence to the identifications.

*Table 2. Overview of DNA datasets (being) harvested by EMODnet Biology*

| Dataset title | EMODnet Catalogue record | Number of occurrences | Type |
|---|---|---|---|
| ARMS-MBON data on long-term monitoring of hard-bottom communities: ITS results from 2018-2020 | https://emodnet.ec.europa.eu/geonetwork/srv/eng/catalog.search#/metadata/6d617269-6e65-696e-666f-000000008612 | 493 | Metabarcoding |
| ARMS-MBON data on long-term monitoring of hard-bottom communities: COI results from 2018-2020 | https://emodnet.ec.europa.eu/geonetwork/srv/eng/catalog.search#/metadata/6d617269-6e65-696e-666f-000000008357 | 19,402 | Metabarcoding |
| ARMS-MBON data on long-term monitoring of hard-bottom communities: 18S results from 2018-2020 | https://emodnet.ec.europa.eu/geonetwork/srv/eng/catalog.search#/metadata/6d617269-6e65-696e-666f-000000008617 | 21,482 | Metabarcoding |
| MONICEPH - Monitoring cephalopods during whale watching activity in the Azores | https://emodnet.ec.europa.eu/geonetwork/srv/eng/catalog.search#/metadata/6d617269-6e65-696e-666f-000000008748 | 182 | Enriched occurrences |

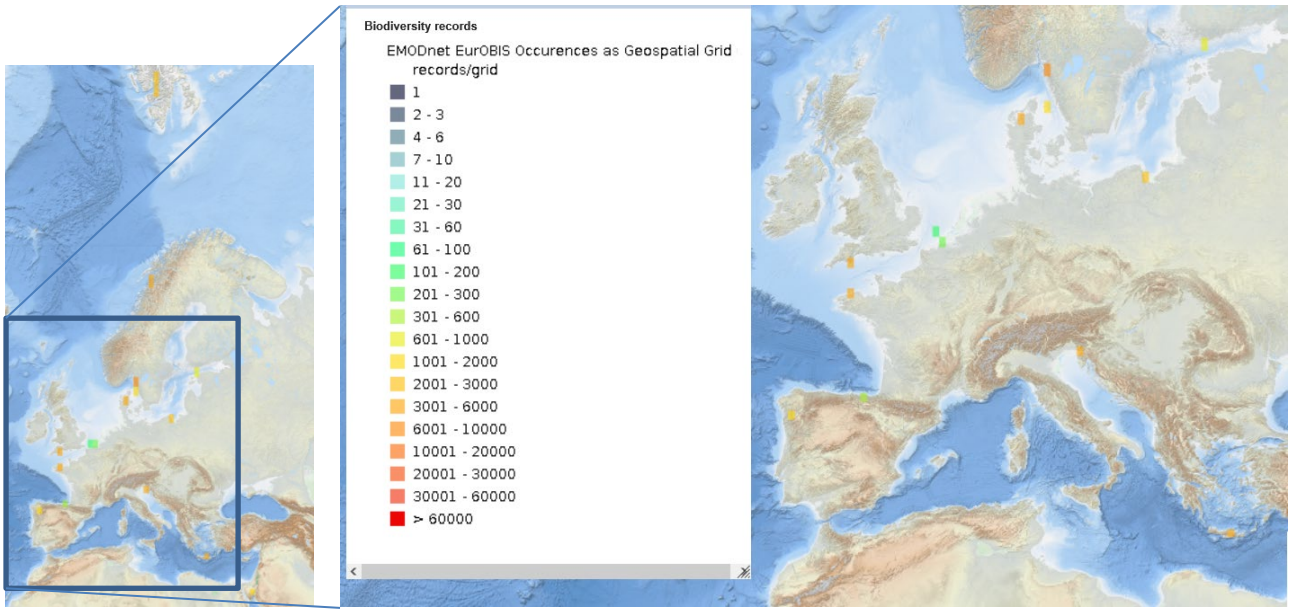



*Figure 2. Map of DNA derived occurrences from the three ARMS datasets published in EMODnet Biology.*