EMODnet

European Marine Observation and Data Network

# EMODnet Biology

**EASME/EMFF/2016/006**

**EMODnet Phase III**

## D2.3: Data standardization and formatting of the datasets mentioned under data coverage section of proposal for linking with EMODnet biology

*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[1]

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 EASME or of the European Commission. Neither the EASME, nor the European Commission, guarantee the accuracy of the data included in this study. Neither the EASME, the European Commission nor any person acting on the EASME'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 | D2.3: Data standardization and formatting of the datasets mentioned under data coverage section of proposal for linking with EMODnet Biology |
| WP title | WP2: Data access to marine biological data |
| Task | Task 1. A common method of access to data held in repositories |
| Authors | Leen Vandepitte (VLIZ), Paula Oset (VLIZ) |
| Dissemination level | Public |

[1] The disclaimer is needed when the document is published

# Contents

# 1 Marine biological data in EMODnet Biology – general assessment

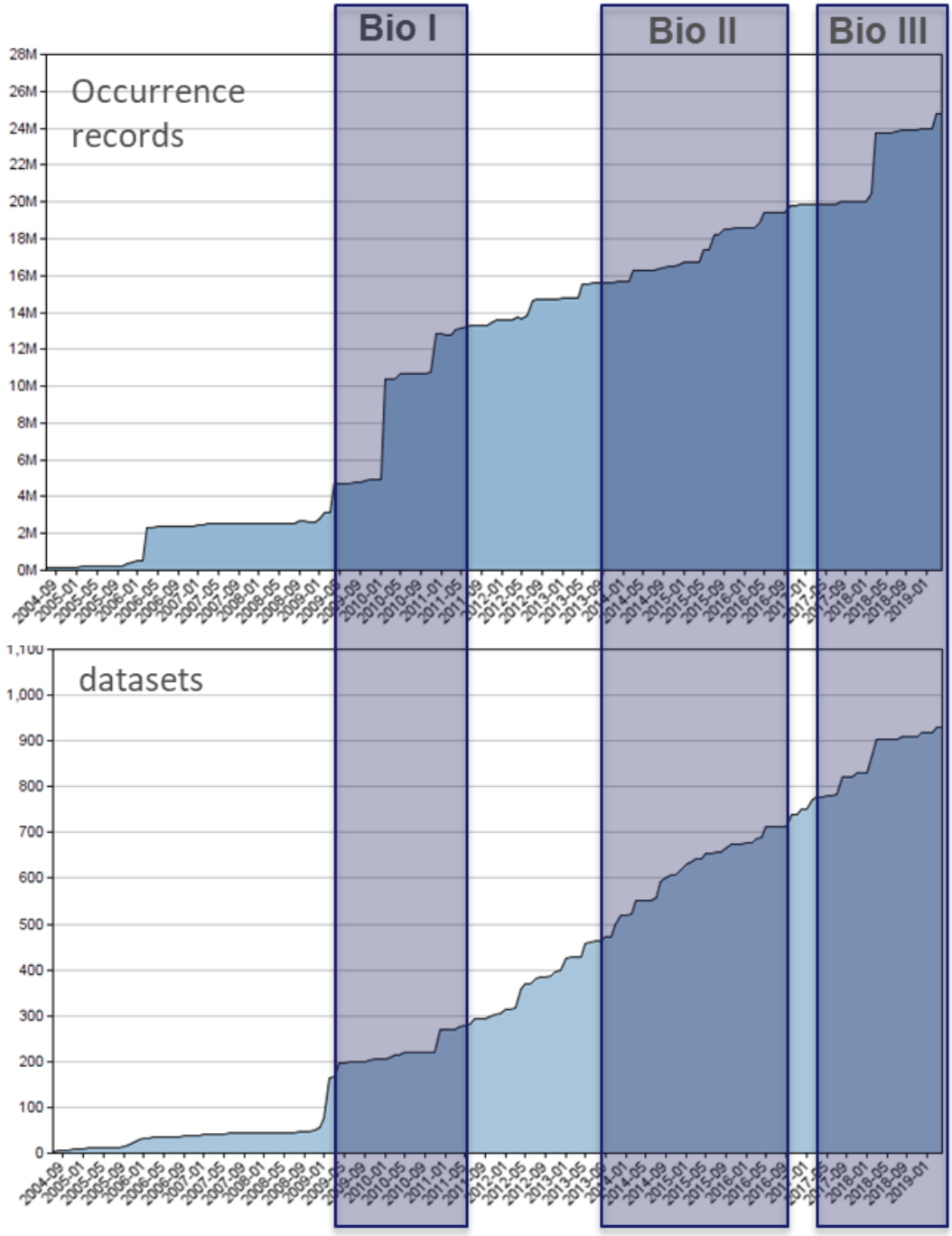
## 1.1 Increase in available datasets and distribution records

On 6 June 2019 – date of report finalization – 874 datasets are available in the EurOBIS database, which is the backbone database for the EMODnet Biology Portal. These datasets represent 25.279.688 occurrence records, of which 19.667.353 (=78%) pass the required quality control procedures to be valid for the EMODnet Biology Portal. For a large portion of these records, extended data – such as abundances and biomass information - is available.

*Table 1. Datasets per partner, linked to D2.2 and D2.3: Data standardization and formatting of the data. Not all of these datasets are yet available online; a large amount of the datasets will be delivered in the next phase of the project, as there were problems with the processing of the dataset (e.g. change in staff at the partner side). (D2.3)* includes the datasets from the first priority.*

| Partner | # new datasets "Priority 1", M12 (D2.2) | All datasets (M0 – M24) (D2.3)* |
|---|---|---|
| Aarhus University (AU) | - | 2 |
| Centre for Environment, Fisheries and Aquaculture Science (CEFAS) | 5 | 5 |
| Deltares | 1 | 2 |
| Finnish Environment Institute (SYKE) | 1 | 3 |
| Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) | - | 1 |
| Institute of Agricultural and Fisheries research (ILVO) | 3 | 4 |
| Institute of Marine Research (IMR) | 3 | 6 |
| Instituto Español de Oceanografía (IEO) | - | 1 |
| Instituto Português do Mar e da Atmosfera (IPMA) | - | 7 |
| International Council for the Exploration of the Sea (ICES) | 5 | 8 |
| Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS) | 8 | 20 |
| Marine Biological Association of the UK (MBA) | 10 | 6 |
| National Institute for Marine Research and Development "Grigore Antipa" (NIMRD) | 3 | 8 |
| Royal Netherlands Institute of Sea Research (NIOZ) | - | 9 |
| Ruđer Bošković Institute (RBI); Center for Marine Research | 2 | 6 |
| Swedish Meteorological and Hydrological Institute (SMHI) | 2 | 5 |
| Sir Alister Hardy Foundation for Ocean Science (SAHFOS) | - | - |
| **TOTAL** | **43** | **93** |

*Figure 1. General overview of the growth in number of records and number of datasets available through EurOBIS, with indication of the specific growths during the different phases of the EMODnet Biology project.*




EurOBIS (from 2014 on) opened up to additional biotic and environmental measurements with the measurement or facts extension to the occurrence table in 2014. For many datasets harvested before 2014, additional information is available and an effort is undertaken include these measurements in EurOBIS as well.

*Figure 2. Evolution in the growth of number of occurrence records in EurOBIS per functional group.*

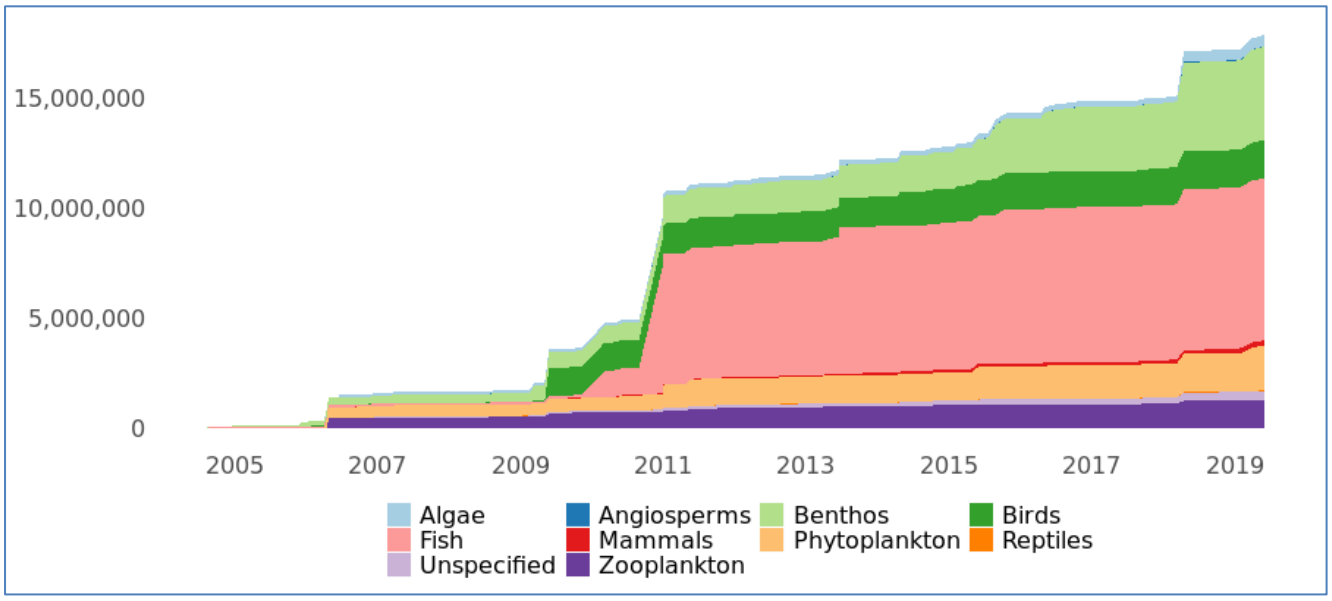
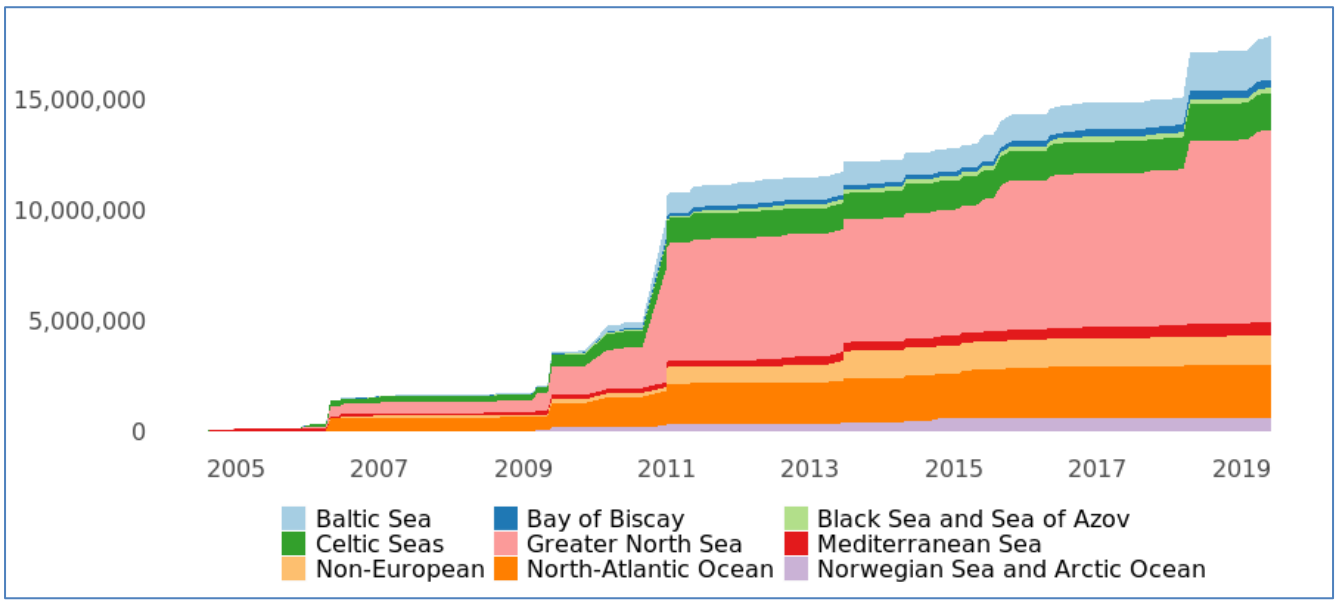



*Figure 3. Evolution in the growth of number of occurrence records in EurOBIS per region.*




## 1.2 Data delivery mechanisms

The majority of the new datasets (68) are delivered through the Integrated Publishing Toolkit (IPT), the preferred data exchange mechanism to integrate new data into the EurOBIS database. A list of the data exchange mechanisms chosen by each partners and the number of datasets delivered with each mechanism is listed in the table below.

*Table 2. Datasets per partner and per data delivery mechanism.*

| Partner | IPT | Excel | Local web services | WFS | BIO-ODV |
|---|---|---|---|---|---|
| Aarhus University (AU) | | | | | 2 |
| Centre for Environment, Fisheries and Aquaculture Science (CEFAS) | | 5 | | | |
| Deltares | | | | 2 | |
| Finnish Environment Institute (SYKE) | 1 | | 2 | | |
| Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER) | | | | | 1 |
| Institute of Agricultural and Fisheries research (ILVO) | 4 | | | | |
| Institute of Marine Research (IMR) | 6 | | | | |
| Instituto Español de Oceanografía (IEO) | | 1 | | | |
| Instituto Português do Mar e da Atmosfera (IPMA) | 7 | | | | |
| International Council for the Exploration of the Sea (ICES) | | | 8 | | |
| Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS) | 20 | | | | |
| Marine Biological Association of the UK (MBA) | 6 | | | | |
| National Institute for Marine Research and Development "Grigore Antipa" (NIMRD) | 8 | | | | |
| Royal Netherlands Institute of Sea Research (NIOZ) | 9 | | | | |
| Ruđer Bošković Institute (RBI); Center for Marine Research | 6 | | | | |
| Swedish Meteorological and Hydrological Institute (SMHI) | 1 | | 4 | | |
| ***TOTAL*** | **68** | **6** | **14** | **2** | **3** |

# 2 Marine biological data in EMODnet Biology – per regional sea

About 1.3 million species occurrence records are located outside the European regional seas discussed in the following section and are therefore excluded from further analyses.

### *2.1.1 Baltic Sea*

EMODnet Biology provides access to 79 datasets which contribute species data for the Baltic Sea, with a total of 1.972.482 species occurrence records. The occurrences are spread between 1754 and 2018 and cover all the functional groups.

The figure and the distribution map provide insight on the temporal and spatial distribution of the available data per functional group in this region.

*Figure 4. Total number of species occurrence records per sample year for the Baltic Sea, subdivided per functional group.*

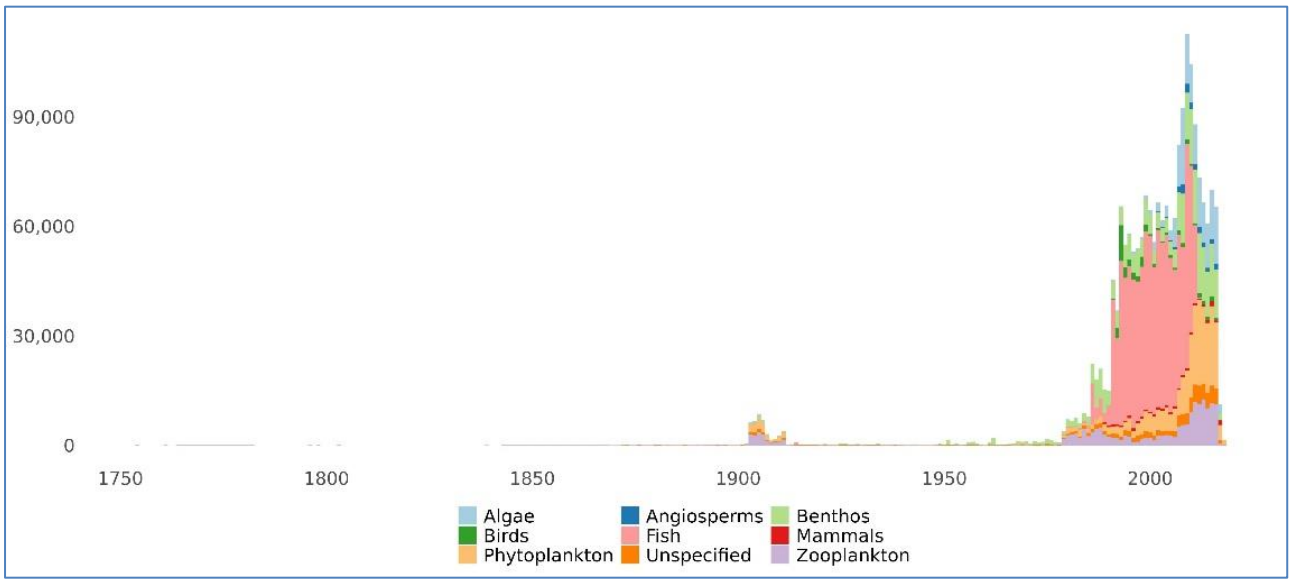
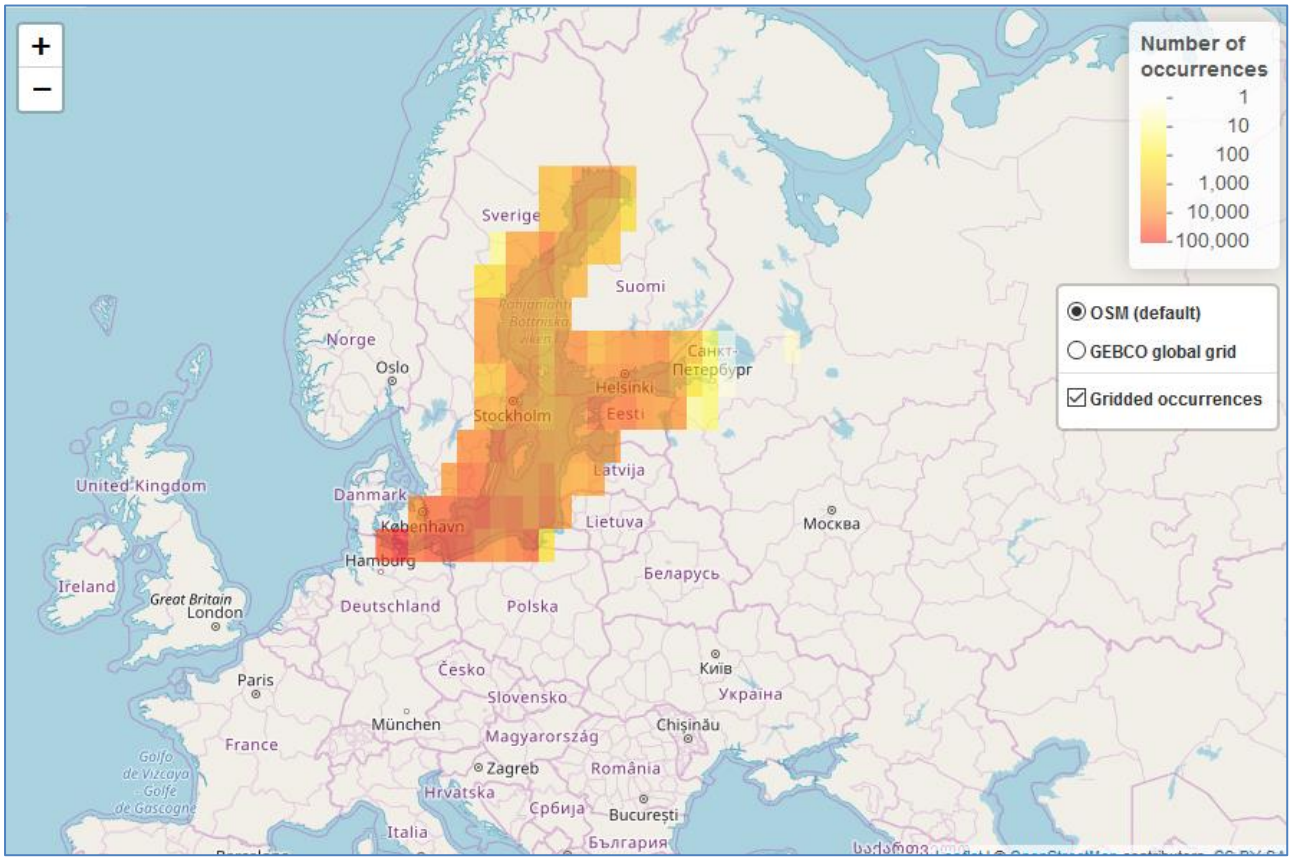



*Figure 5. Gridded distribution of occurrence records in the Baltic Sea.*



### *2.1.2 Bay of Biscay*

EMODnet Biology provides access to 88 datasets which contribute data for the Bay of Biscay, Iberian Coast and Macaronesia, with a total of 359.207 species occurrence records.

The figure and the distribution map provide insight on the temporal and spatial distribution of the available data per functional group in this region.

*Figure 6. Total number of species occurrence records per sample year for the Bay of Biscay, Iberian Coast and Macaronesia, subdivided per functional group.*

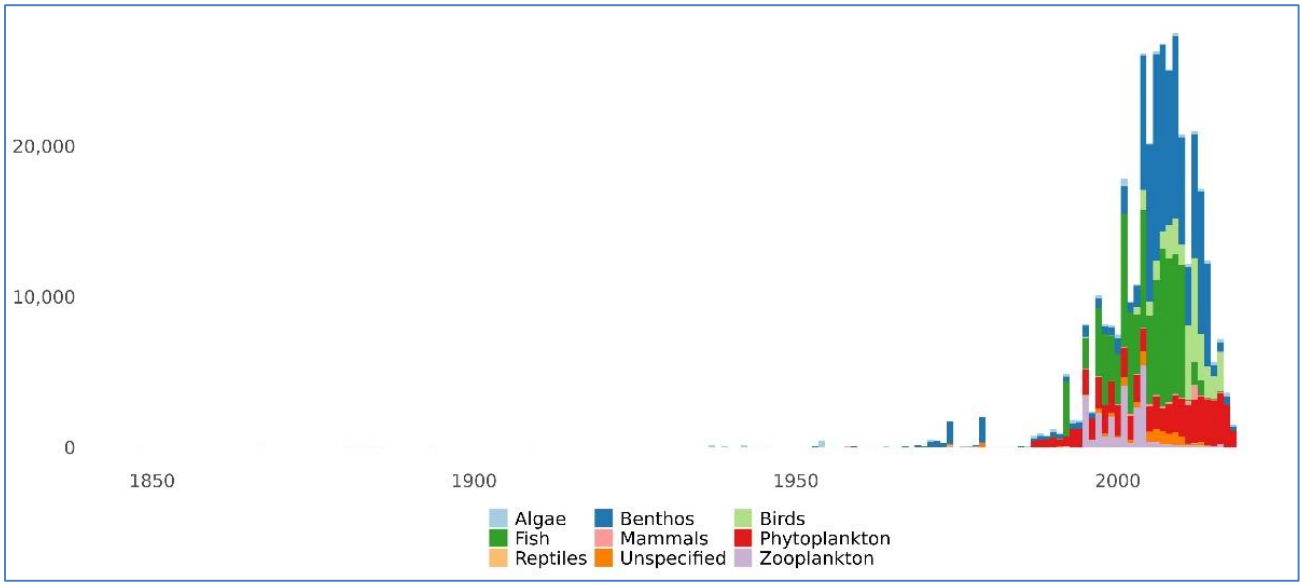
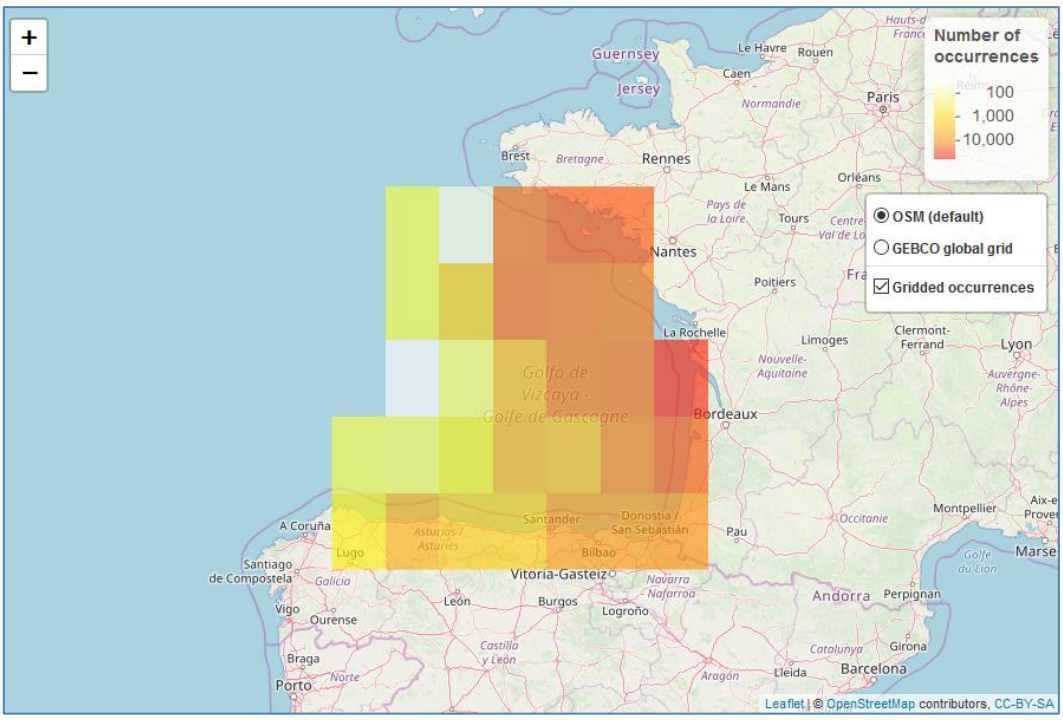



*Figure 7. Gridded distribution of occurrence records in the Bay of Biscay.*



### *2.1.3 Black Sea and Sea of Azov*

EMODnet Biology provides access to 105 datasets which contribute data for the Black Sea and Sea of Azov with a total of 245,138 species occurrence records.

The figure and the distribution map for this region provide insight on the temporal and spatial distribution of the available data per functional group in this region.

*Figure 8. Total number of species occurrence records per sample year for the Black Sea and Sea of Azov, subdivided per functional group.*

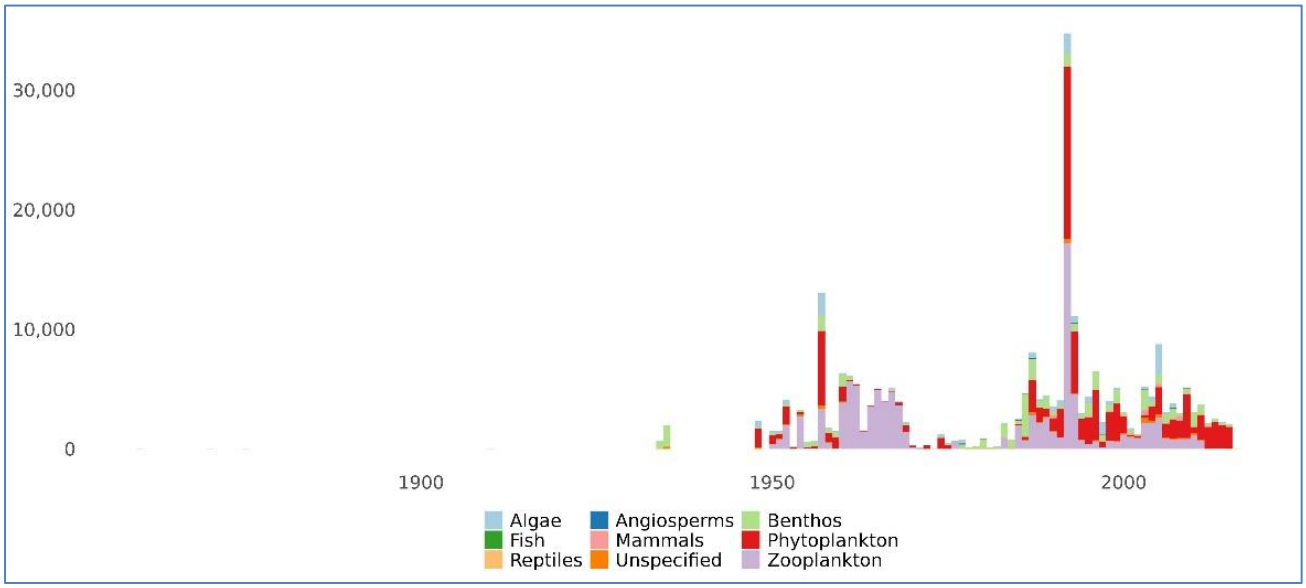
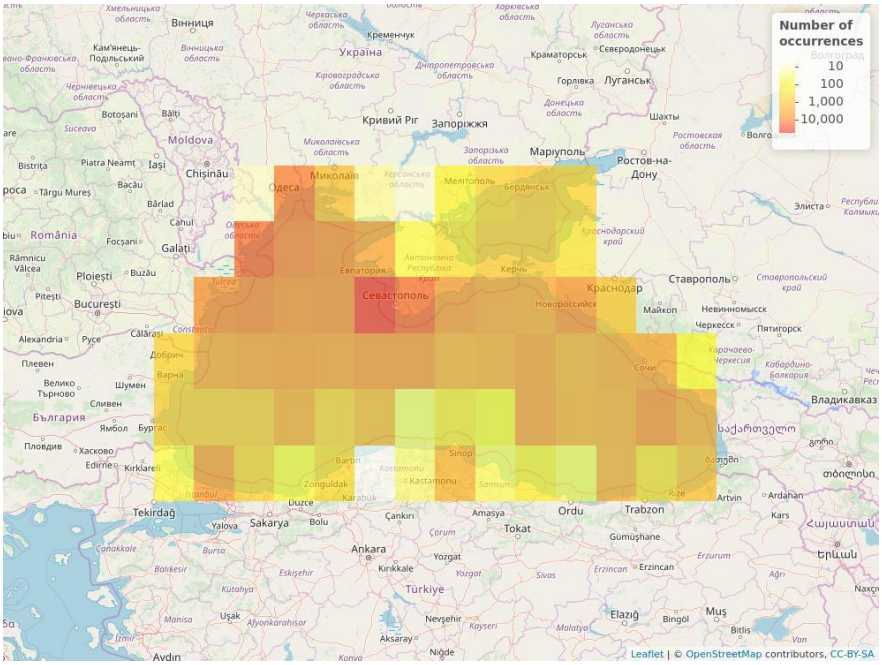



*Figure 9. Gridded distribution of occurrence records in the Black Sea and Sea of Azov.*



### 2.1.4 Greater North Sea

EMODnet Biology provides access to 284 datasets which contribute data for the Greater North Sea with a total of 8,673,577 species occurrence records.

The figure and the distribution map for this region provide insight on the temporal and spatial distribution of the available data per functional group in this region.

*Figure 10. Total number of species occurrence records per sample year for the Greater North Sea, subdivided per functional group.*

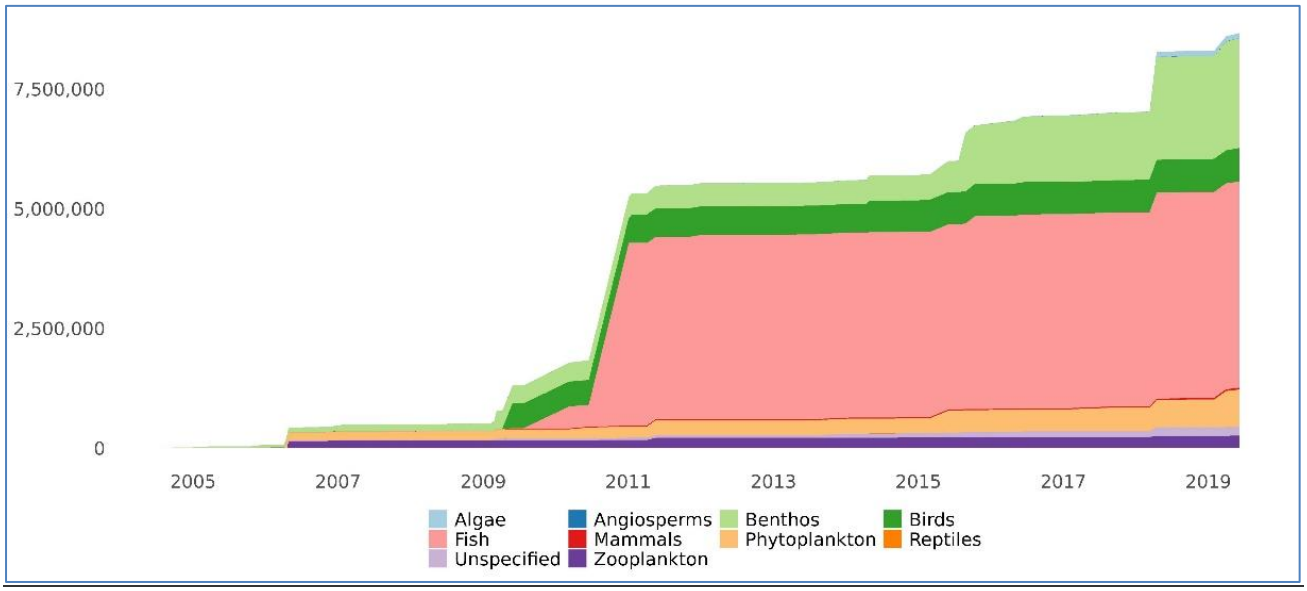
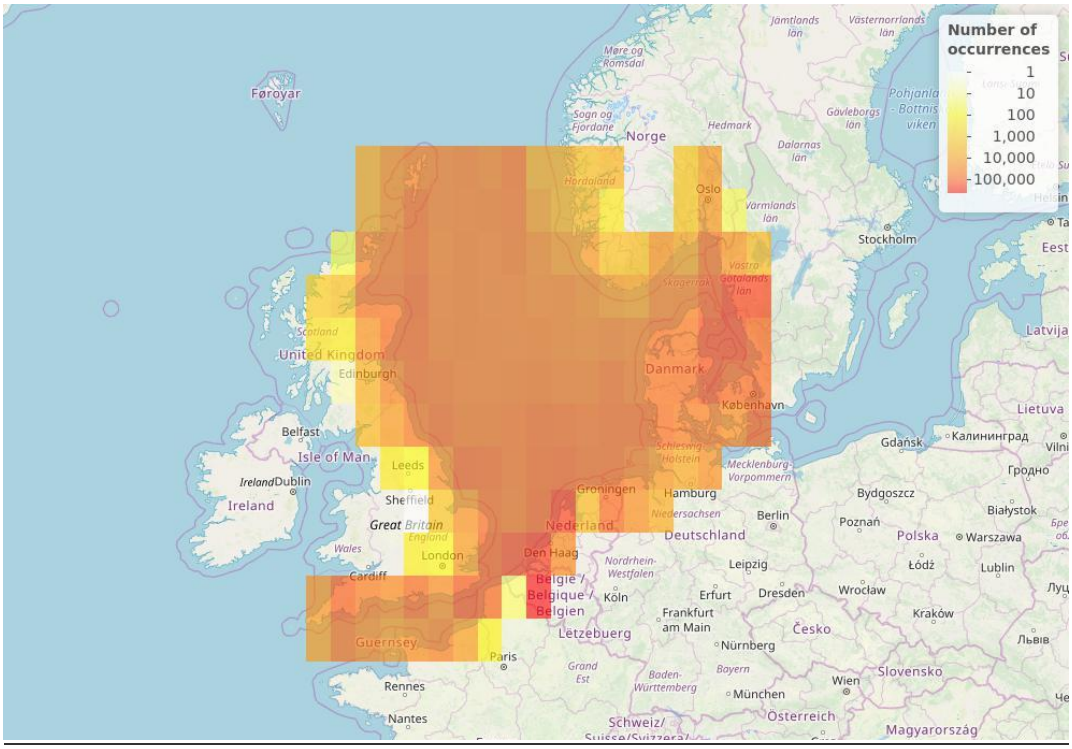



*Figure 11. Gridded distribution of occurrence records in the Greater North Sea.*



### 2.1.5 Celtic Seas

EMODnet Biology provides access to 116 datasets which contribute data for the Celtic Seas with a total of 1,666,199 species occurrence records.

The figure and the distribution map for this region provide insight on the temporal and spatial distribution of the available data per functional group in this region.

*Figure 12. Total number of species occurrence records per sample year for the Celtic Seas, subdivided per functional group.*

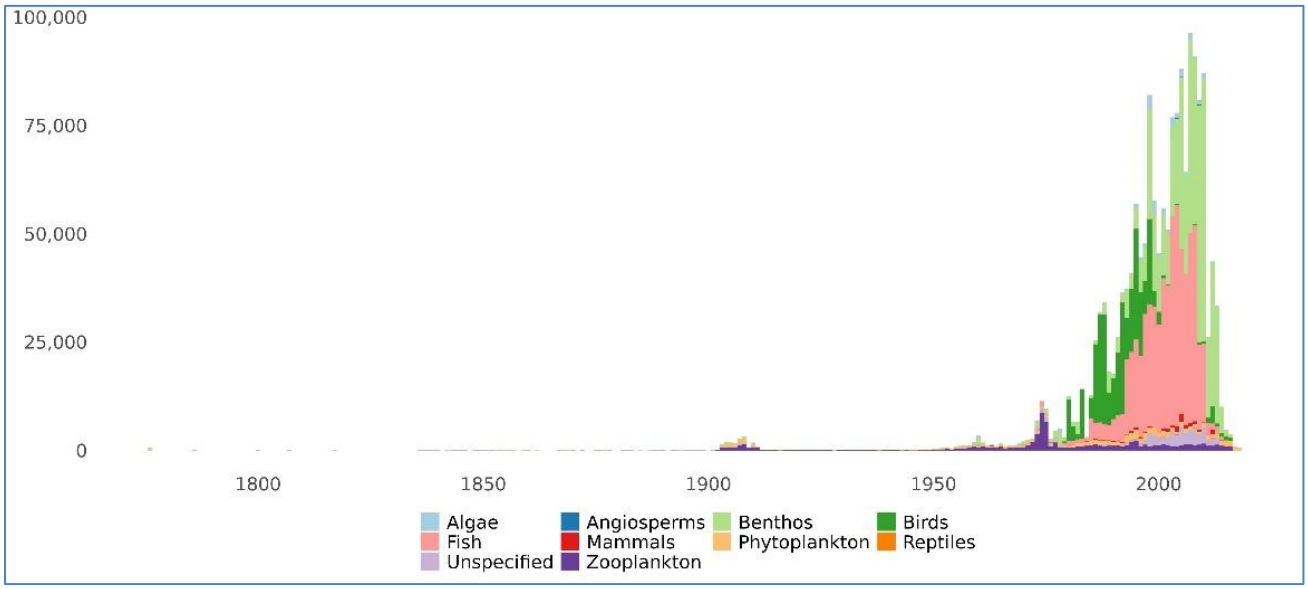
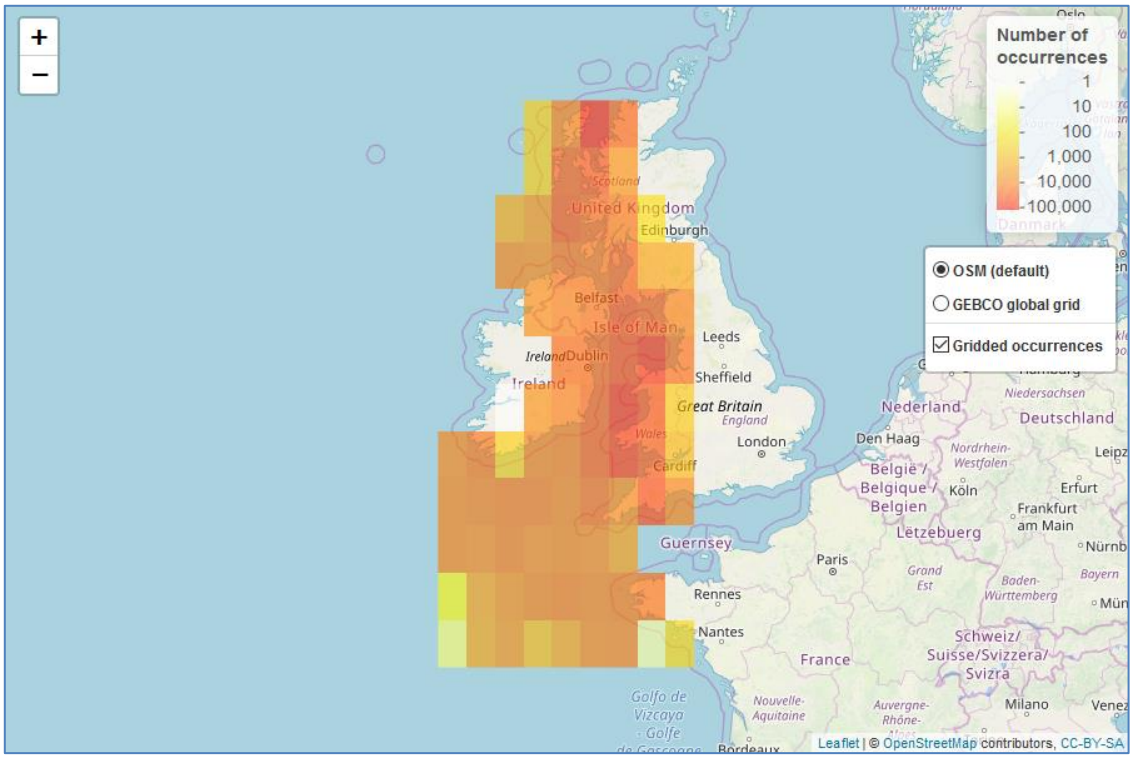



*Figure 13. Gridded distribution of occurrence records in the Celtic Seas.*



### 2.1.6 Mediterranean Sea

EMODnet Biology provides access to 212 datasets which contribute data for the Mediterranean Sea with a total of 635,034 species occurrence records. We observe thus a relative high number of small datasets.

The figure and the distribution map provide insight on the temporal and spatial distribution of the available data per functional group in this region.

*Figure 14. Total number of species occurrence records per sample year for the Celtic Seas, subdivided per functional group.*

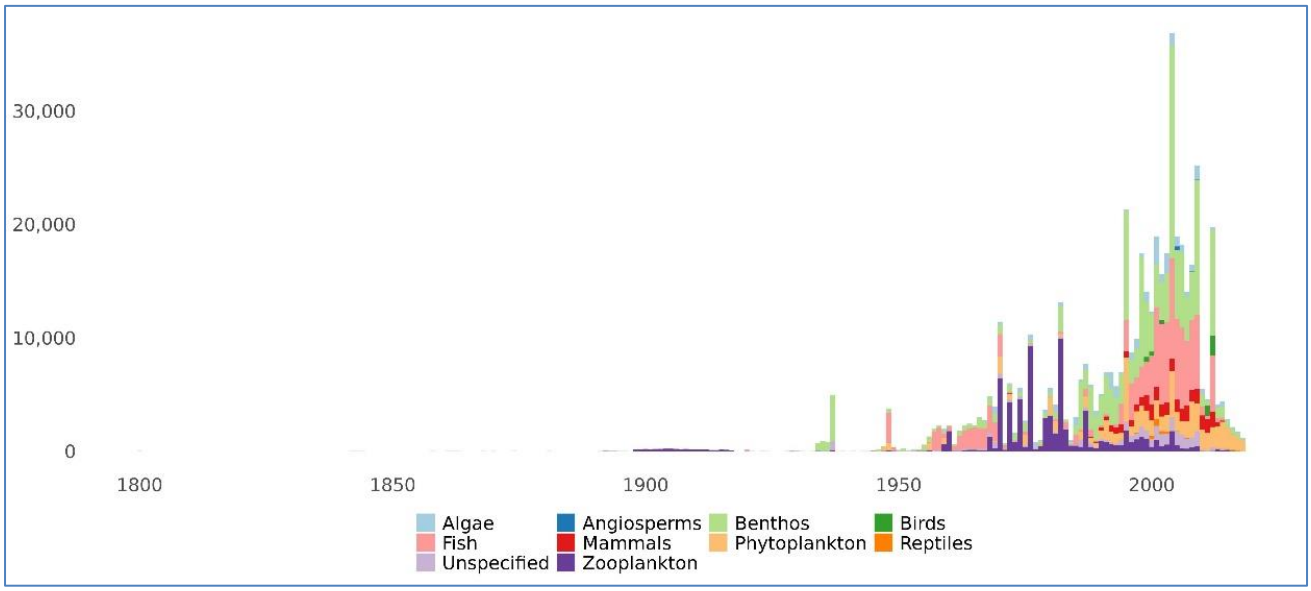
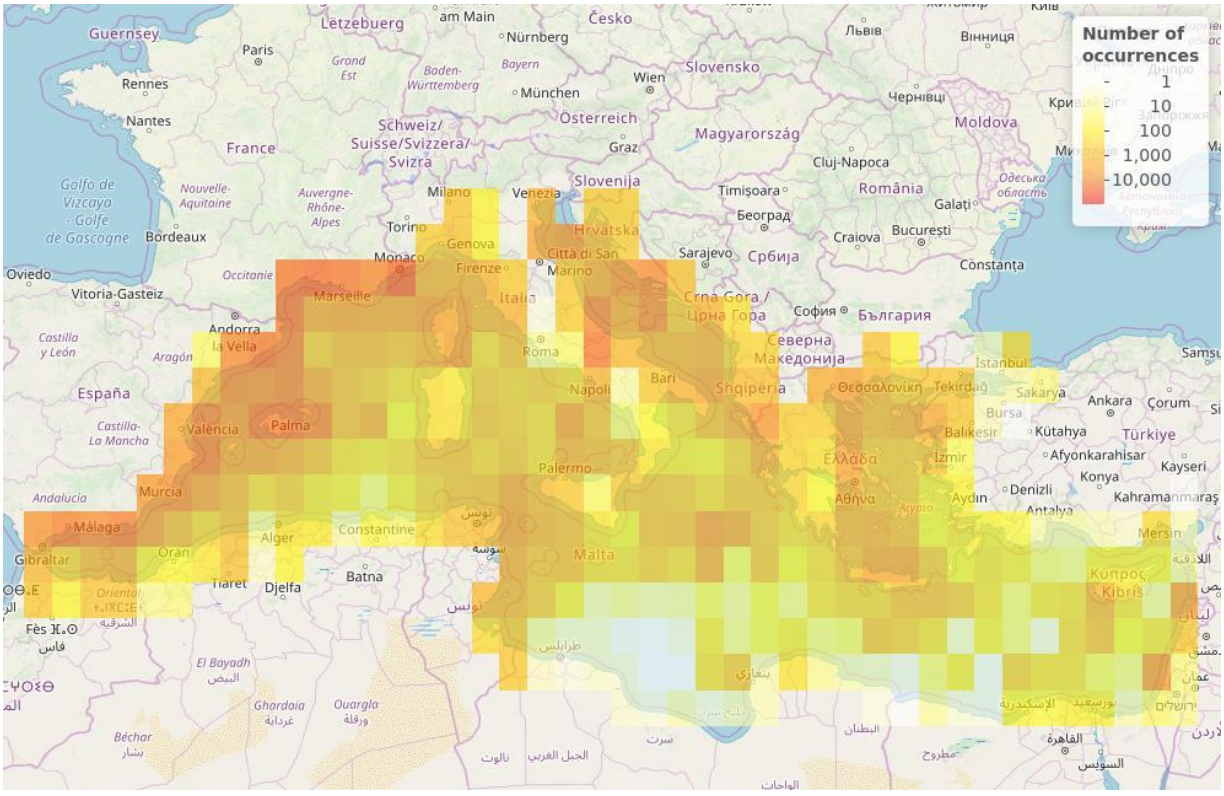



*Figure 15. Gridded distribution of occurrence records in the Mediterranean Sea.*



### 2.1.7 Norwegian Sea & Arctic Ocean

EMODnet Biology provides access to 134 datasets which contribute data for the Norwegian Sea & Arctic Ocean with a total of 657,899 species occurrence records. The chart and maps provide insight on the temporal and spatial distribution of the available data per functional group in this region.

*Figure 16. Total number of species occurrence records per sample year for the Norwegian Sea & Arctic Ocean, subdivided per functional group.*

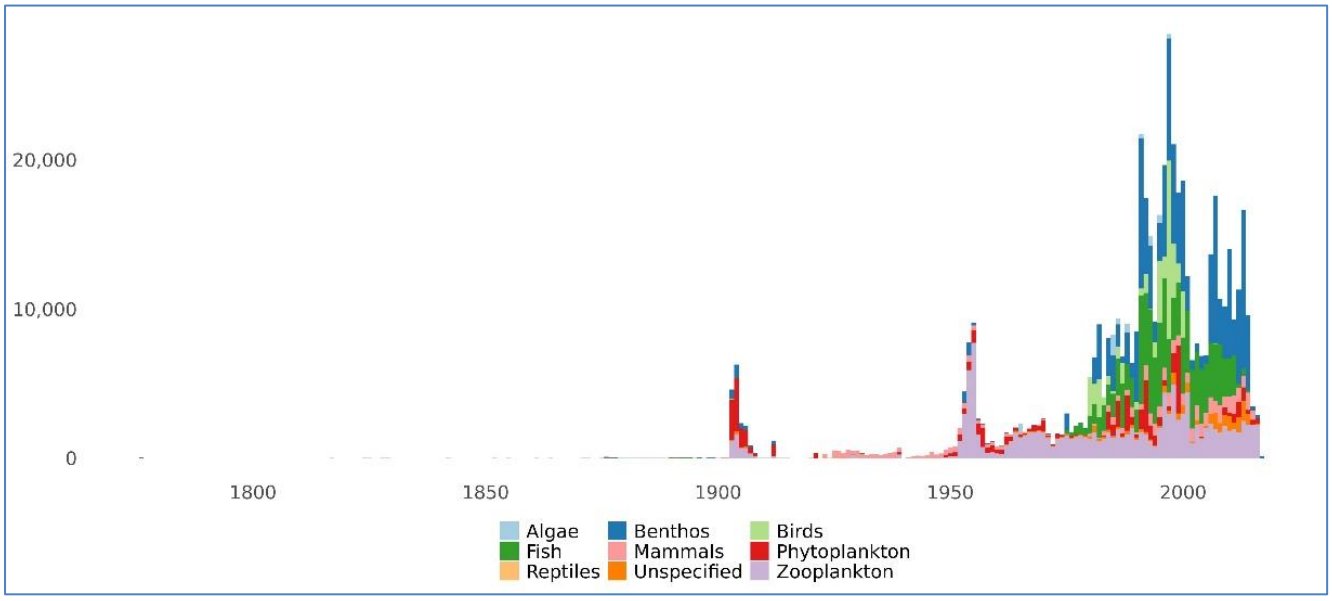
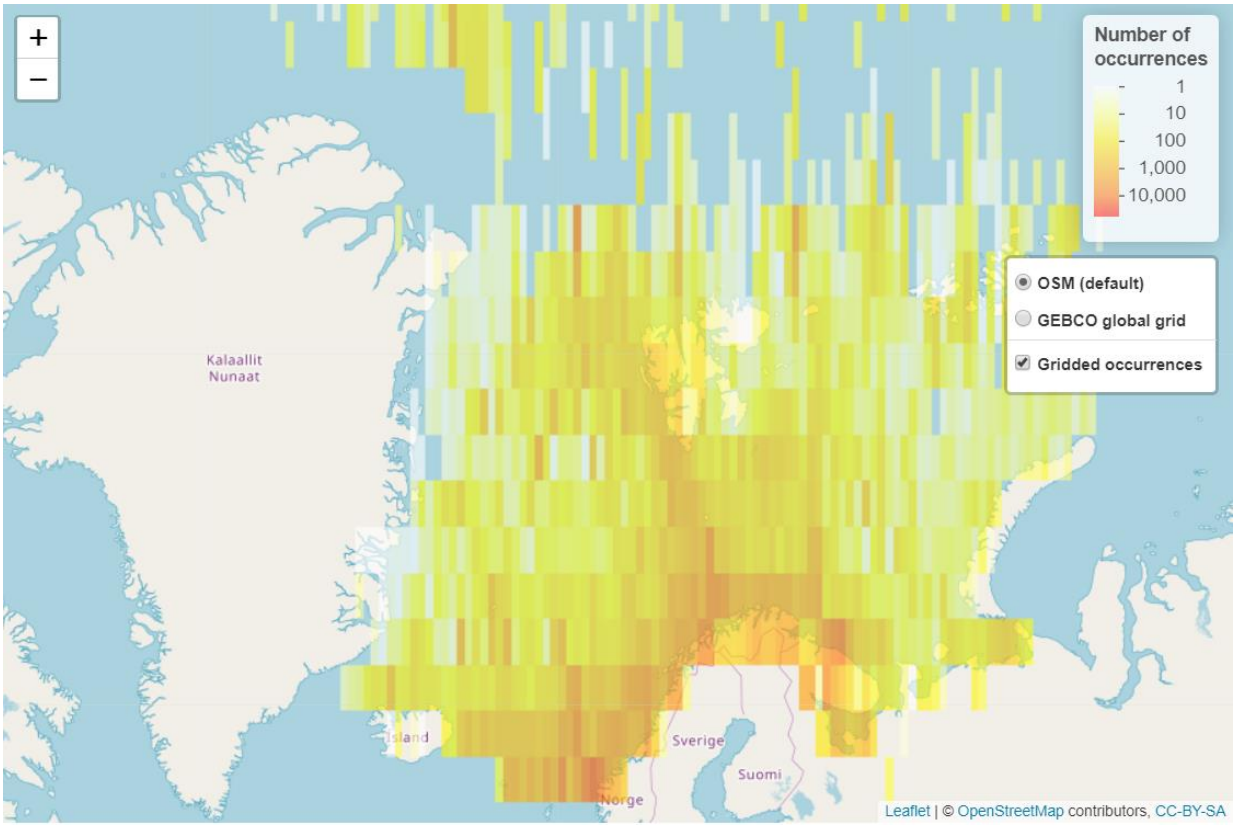



*Figure 17. Gridded distribution of occurrence records in the Norwegian Sea and Arctic Ocean.*



# 3 Data standardization and quality checking

All datasets are either formatted according to the data model followed by EMODnet Biology and (Eur)OBIS. This data formatting implies transformation from the original files into DwC Event or Occurrence core. All the fields present in the original data are mapped to DwC terms and the content of the data is formatted according to specific standards: ISO_8601 for date, World Register of Marine Species LSID for taxonomy, BODC vocabularies for the data in the Measurements or Facts extension (in Event core). Once the data is formatted, a quality check is performed. This QC are performed automatically but the online QC tool, and the following tests are involved:

- Dataset integrity:
    - 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 link the same one as in the occurrence extension?
    - Are there 'duplicate occurrences' meaning is the same taxon listed twice at the same EventID without any difference in any of the biometric parameters?
    - Are there 'duplicate measurements' meaning does the same measurement occur twice for the same occurrenceID or the same EventID?
- Are all mandatory fields present in the dataset?
- Are all mandatory field filled out?
- Does eventDate follow the required format?
- Does the scientificNameID 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?
- Does the measurementTypeID / measurementValueID refer to an existing term in the BODC vocabulary?
- 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
- Is there a sampling instrument present?
- Are there other sampling descriptors present?
- Provide an overview of the number of taxa per kingdom and class.
- List the non-matched taxa (including deleted and quarantined matches)
- Plots of the coordinates and the distribution of the temporal cover are provided to allow for quick comparison with the metadata.