EMODnet Biology

EASME/EMFF/2016/0066

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

EMODnet Phase III

D4.5: Proposed list of data packages and underlying datasets with specific recommendations on (meta)data gaps

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

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 [ref]*	Report with proposed list of data packages and underlying datasets with specific recommendations on (meta)data gaps
WP title [ref]*	WP4: Data product creation
Task [ref]*	T1: Critical assessment of metadata and data gaps
Authors [affiliation]	Peter M.J. Herman [Deltares], Lennert Schepers [VLIZ], Joana Beja [VLIZ], Leen Vandepitte [VLIZ], Charles Troupin [ULg], Alexander Barth [ULg], Tom Webb [USheffield], Karline Soetaert [NIOZ], Dan Lear [MBA], Silvana Birchenough [CEFAS]
Dissemination level	Public
Submission date	18/11/2019
Deliverable due date	M5

---

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

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

Contents

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

Recommendations on (meta)data gaps

1 Introduction

For the production of the Atlas of Marine Life, EMODnet Biology has decided to follow the general structure of the Essential Ocean Variables (EOVs). This entails that information is to be gathered on such variables as phytoplankton species composition, macrozoobenthos density, biomass and species composition etc. A prerequisite for the production of gridded maps of these variables is that a consistent compilation of available data relating to the variable can easily be made based on the database. The procedure to be followed is usually the following sequence: select all useful data from the database, check for methodological and taxonomic consistency and correct where needed (e.g. lump taxonomic groups: in cases where one dataset distinguishes subspecies and the other doesn't, all data are lumped to species level), find appropriate normalisation for the sampling gears used, combine datasets and produce maps.

In order to automate the first steps of this procedure, there is a need to quickly and automatically select datasets that contribute to the knowledge of the different EOVs in a pre-defined geographical area. Dataset metadata are essential for this selection procedure, and the present document proposes to extend these metadata, largely on the basis of knowledge already contained in the database.

In addition, the preparation of distribution maps of single species is hampered by the ambiguity of absence information on a species. Presence of a species in a grid cell is certain if it has been observed in a sample within that grid cell. However, if a species has never been recorded in a grid cell, this does not necessarily indicate it is absent. An obvious reason for non-recording is that not a single sample capable of finding the species has been included in the database. In this case there is no information on the presence or absence of the species. However, if samples of a suitable type have been taken but the species was never recorded, the observed species density can be set to 0. This is not necessarily the 'true' value (the species can be present but due to chance be missing in all samples taken) but it is the best possible estimate given the sampling effort. For the sake of gridding species abundance or species presence, the inclusion of 'non-available' (case 1 of no samples) or of the numerical value 0 (case 2 of samples without the species) makes a large difference.

The construction of overview maps on the coverage of EOVs, as well as the preparation of statistically sound interpolated maps for species, can be greatly enhanced by maximally exploiting and/or complementing the metadata associated to the records in the database. At different levels in the database (e.g. species, data sets, methods, functional groups) information is available or can be made available that will allow producing the desired information. Much of that information can be obtained by data mining, but manual input and editing may be needed at crucial points.

The purpose of this document is to identify the needs for data mining and expert checks in order to optimize the use of the EMODnet Biology database. We start the discussion with an example and generalize from there towards the general problem. The document ends with recommendations for future development of the database.

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

2 Example problem 1: diversity of macrobenthos in the Southern Bight of the North Sea

In this analysis, we attempted to provide an overview of the information in EMODnet Biology that is suitable to estimate biodiversity hotspots in the macrobenthos of the Southern Bight. This would provide crucial information for marine spatial planning, environmental impact studies etc. The exercise is discussed in Lear et al. (subm.). Here we provide the main results in Figure 1, together with a summary of the methodology.

Figure 1A summarizes the desired outcome of the exercise. The biodiversity of the macrobenthos was here simply approached as the average (per grid cell) number of species per sample, where a 'sample' was identified as any sampling event with common time and place catching at least two different species. It can be deduced from Fig. 1A that spatial patterns in this measure of diversity exist. Specific areas along the UK east coast, as well as inner areas in the Southern Bight (Frisian Front, Oyster grounds) show higher than average diversity. Shallow areas with dynamic sands are markedly poorer in species. The figure also illustrates that at the relatively fine spatial grain (approximately 5*5 km^2) overall sampling of the North Sea is scarce. This is further illustrated by Fig. 1B, showing that at this resolution approximately 2/3 of all grid cells have never been sampled (or the samples haven't made it into the EurOBIS database). Sampling effort is very unevenly spread over space (Fig. 1C), with disproportionate efforts in shallow coastal areas. Fortunately, however, most of the Southern Bight is covered with relatively recent samples (Fig. 1D).

These results were obtained using the following procedure:

- Download all EurOBIS data from the geographical area of the Southern Bight
- Download the species attributes of all AphiaIDs in this dataset
- Filter all species on the traits related to being benthos (e.g. some species are characterized as benthos, others as macrobenthos, others have AMBI scores. At present many species in WoRMS have attributes, but not all)
- List all datasets to which the occurrences of benthic species belong
- Filter manually the datasets: eliminate museum collections, epifauna collections, zooplankton and fishery cruises, and known incomplete sampling programmes (e.g. shellfish surveys)
- From the original download of all records, only keep the selected datasets
- Find sampling events based on time and place, calculate number of species per sampling event and other statistics
- Per dataset, calculate the total number of sampling events, and the total number of species found in the dataset. Plot these numbers against each other for all datasets on log-log plot. Identify those datasets that clearly fall out of the relation as potentially 'incomplete', i.e. not sampling the entire community. Remove these incomplete datasets if, from the descriptive metadata or the complete species list in the dataset, the incompleteness can be confirmed.
- Repeat the calculation of number of species per sampling event in the purified list of datasets
- Grid and map

The main problems encountered in this analysis are discussed below.

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

A surprising observation was that more records containing information on truly benthic species were discarded than used. This is caused by the fact that many zooplankton samples (notably the Continuous Plankton Recorder [CPR] data) contain presence information of larvae of benthic animals. While this may be valuable data with respect to recruitment processes in the benthos, it does not yield information relevant for benthic community diversity studies. Automatic filtering of datasets would thus require knowledge on the EOV the dataset primarily contributes to (in the case of CPR this would be zooplankton species composition), but also on what other EOVs it provides information on (for CPR this would be benthic recruitment, or benthic larvae occurrence).

A second observation was that the trait information in WoRMS, e.g. specifying whether a species is benthic, is extensive but not entirely complete. With the procedure followed, we retained 970000 records. 860000 of these were records of species called 'benthic' in WoRMS. It is possible that the remainder were mainly species of other groups occasionally recorded in benthic samples (e.g. juvenile fish). This list most probably requires manual expert checking in order to be finalized.

A particular problem was posed by datasets specifically targeting a subgroup of the macrobenthos. A few large datasets were excluded because they biased the estimates of biodiversity of macrobenthic samples. A Dutch monitoring programme targeting only shellfish and a few large benthic species is an example, as is a large German dataset in the German Bight targeting few species only. Other excluded datasets targeted epibenthos or hyperbenthos. What most (but not all) of these datasets have in common is that they employ other sampling gear than grabs or box cores: hyperbenthic sledge, beam trawl, endobenthic suction dredge. A good vocabulary of sampling methods and indication, per data set, of the sampling method used could help greatly in selecting datasets automatically. We propose to use the BODC vocabulary for this, as this would avoid redundancy in an international perspective.

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
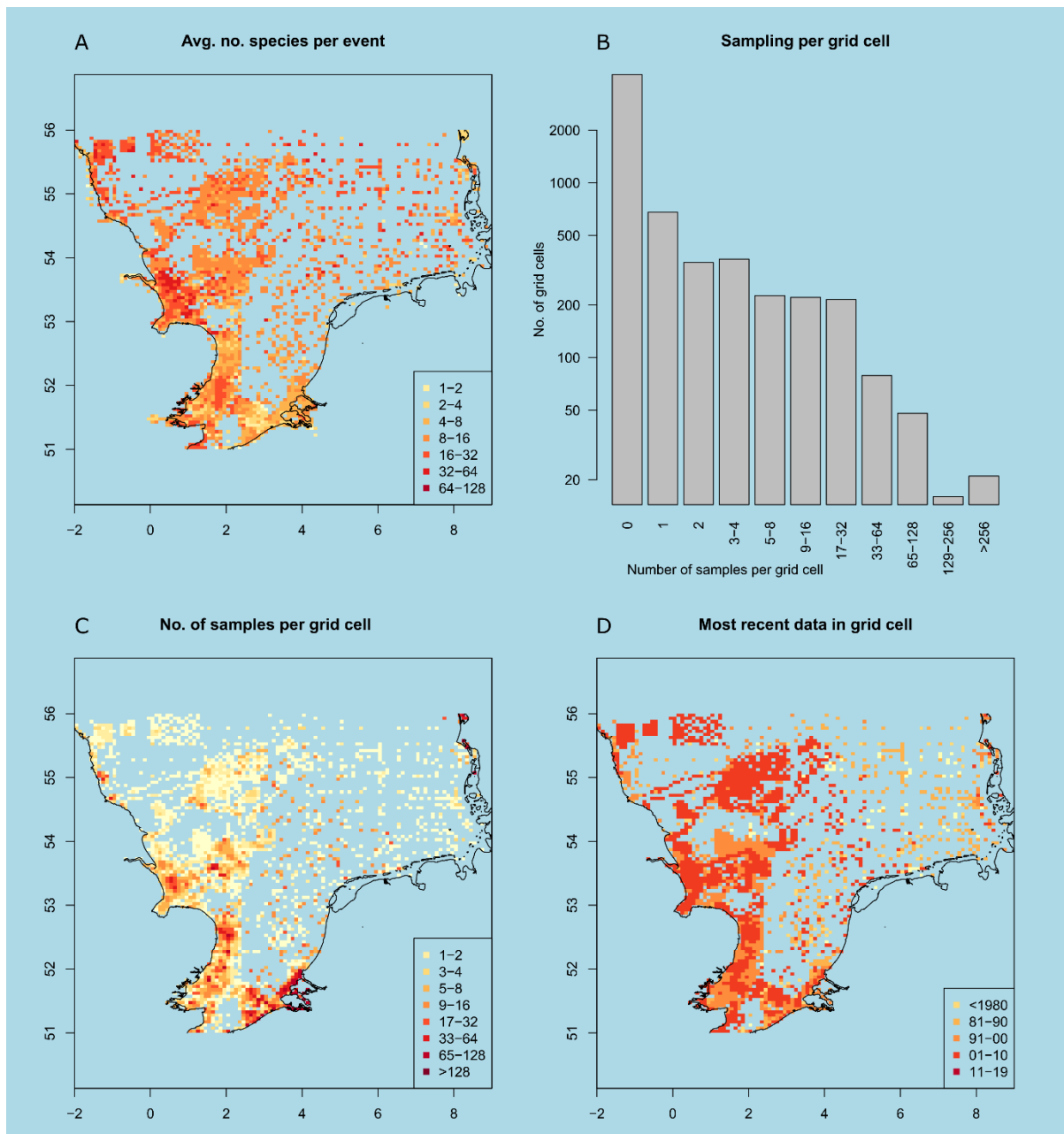with specific recommendations on (meta)data gaps

Figure 1: Overview of data in EurOBIS containing information on benthic macrobenthic community composition. From all records within the polygon, only datasets containing species known to be benthic were retained. Datasets focusing on plankton, fish, epibenthos or meiobenthos were manually removed, as were museum collections containing only a single specimen per sampling event. Sampling events were defined as all species records sharing date and location. These events were gridded in grid cells of 0.1 degree longitude by 0.05 degree latitude, which is order(5 km) in both directions. A. Number of sampling events per grid cell. B. Distribution of number of cells over the classes of sampling frequency. C. Diversity, expressed as the average number of species per sampling event in each of the sampled grid cells. D. Date of the most recent sample per grid cell, showing a clear west-east gradient in how up-to-date the information is.

The distinction between macrobenthos and meiobenthos is also problematic. Some datasets cover both groups, and the information in WoRMS on the specific size class is incomplete. By

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
*D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps*

carefully selecting datasets covering meiobenthic species only and recording the species in these datasets, candidate lists of meiobenthos can be produced and transferred to WoRMS. This will, however, require expert control in order to avoid the misclassification of macrobenthic species occasionally recorded in meiobenthos samples.

When collecting information relevant to estimate community diversity, museum collections interfere with the process. Usually, these collections contain information on isolated species but not necessarily on all species found in the same sample. These datasets are still very valuable to record all confirmed occurrences of a species, but should be indicated separately as they do not contain information on communities or implicit information on the (non)occurrence of other species.

3 Example problem 2: producing maps of species including confirmed absences

In this example, we retrieved the occurrence pattern of the mollusc species *Donax vittatus*, a species known to occur in dynamic sand habitats. Most of its habitat is concentrated near sandy coasts, but it also occurs in wave-swept shallow offshore sites, e.g. at the top of sand banks. Using the same data collection, we prepared the following datasets:

- All presences recorded in the datasets that provide complete sampling of the macrobenthic community. These are the datasets used in the previous exercise, as completeness of the species record was important for the purpose of diversity estimation
- All presences recorded in the complete database, with the exclusion of plankton datasets that most probably only record larval stages
- All 'true zero' observations, that are given by the locations of all samples that were targeting *Donax vittatus*, but have not found it. Distinction is made between two different types of samplings targeting the species. When a dataset targets all macroendobenthos, it automatically also targets this species. Thus all records from 'complete' datasets have been used. However, also 'incomplete' datasets, targeting only part of the macrobenthic community, can be useful if the dataset had *Donax vittatus* in its 'target list'. We do not know the target lists of all datasets explicitly, but derived them from the species list in the data set. In our example, *Donax vittatus* appeared to be targeted by the Dutch shellfish survey and the dataset in the German Bight, which added much presence/absence information on the species.

Based on these datasets we could present several maps of the distribution of the species (Figure 2). Fig. 2A shows all presences in the complete database, but does not contain any information on absence. Fig. 2B shows presence/absence information (expressed as probability of occurrence per grid cell) based only on datasets targeting the complete macroendobenthic community (same datasets as in exercise 1). Fig. 2C shows all presence/absence information from both the datasets targeting the complete community and the datasets only targeting part of the community but with *Donas vittatus* in their target list. The major different with Fig. 2B is the Dutch shellfish dataset. Fig. 2D shows a kriging interpolation of the data in Fig. 2C, and is presently our best approximation of a gridded map of the species occurrence in the Southern Bight of the North Sea. Note that further improvement could be reached by using the

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

presence/absence information from Fig. 2C in a Species Distribution Model, employing additional information from environmental variables. This has not been pursued in the present exercise.
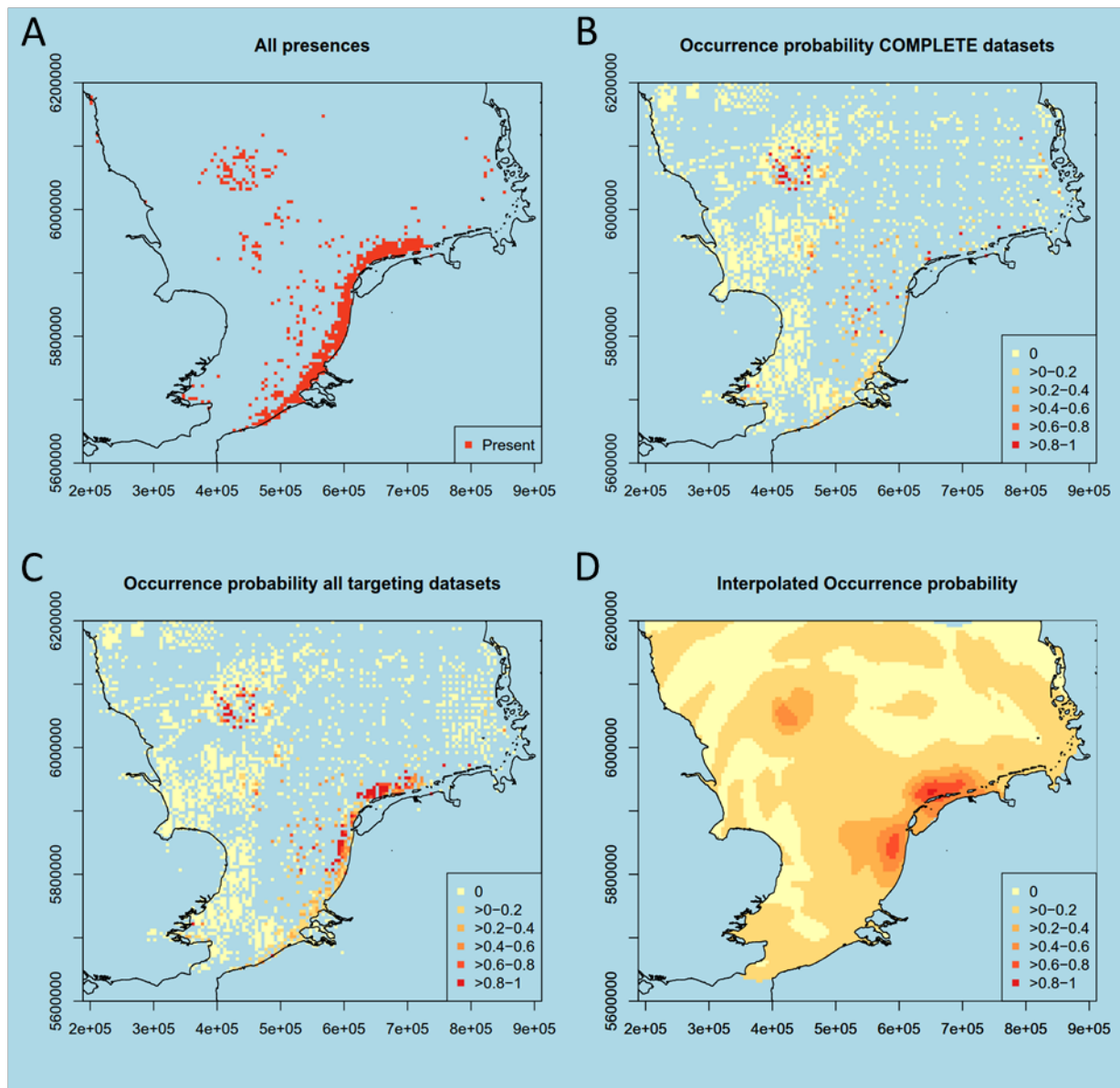

Figure 2. Maps of the occurrence of the mollusc Donax vittatus in the EMODnet Biology database. A. all occurrences in the database, with the exclusion of plankton-oriented datasets that only contain information on larvae. B. occurrences recorded in samples that target the complete macrobenthic community. Occurrence probability has been calculated per grid cell. C. Occurrence probability based on the data sets of (B) plus all incomplete datasets targeting the species.. D. Kriging interpolation based on the occurrences as summarized in (C).

In order to produce these maps, essentially the same information as discussed in Example 1 was needed. It is necessary to know for each dataset what it exactly targets. This does not only relate to whether benthos, plankton or fish is the main target, but also whether the dataset gives complete overviews of the endomacrobenthic community, or only a partial view due to selectivity of the gear or the identification process.

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

The complete distribution maps, including observed zeroes, provides a major step forward for the exploitation of the EMODnet Biology database. Further validation of the zeroes may still be needed, but if we can provide these for a large number of species, this will open up the possibility to provide reliable gridded maps, as well as improved Species Distribution Models, for a large range of species. Completing the database with the essential metadata needed for this type of models will thus provide extremely valuable information to improve the products of the Atlas of Marine Life.

4 Proposal for semi-automatic completion of metadata in the databases

The production of overviews of available information and the production of consistent datasets on (sub)groups of fauna and flora, requires addition of metadata at different levels in the database. At least part of this information is available from diverse sources and can be collected semi-automatically. Here we discuss what meta-information can be added to the database and what are possible sources of the information.

4.1 Species

Species are documented in WoRMS, where also trait information is stored. For the purposes discussed here we propose to add/edit information on:

- The subgroup, or functional type, the species belongs to. This should be specific enough, and thus go beyond the benthic/pelagic distinction. For the functional group 'benthos' we propose sub distinctions based on size (mega, macro, meio, micro), and vertical sub habitat (endo, epi, hyperbenthos). Elaborations of other groups (e.g. zooplankton, phytoplankton, macroalgae, fish) are being made based on the sampling methods and contents of the database. Appendix I gives an overview of the functional types that are presently being implemented in WoRMS.
- Other traits, following the list of priority traits determined by the WoRMS community, are being added to WoRMS already. We propose to include the life history traits of fish and benthos, used in existing data products, to this list. A sub-database of these traits is already available

By screening dataset information, using that information to derive the functional groups species belong to, and by combining several databases, proposed lists of species traits to be added to WoRMS are currently being prepared. Appendix B specifies the details. This work is done by the curators of WoRMS at VLIZ. Final incorporation of this information into WoRMS will take some time, as the taxonomic editors will check the information and have to approve it before it is included in the public data base.

4.2 Datasets

Datasets are much less numerous than individual species occurrence records and provide a good level to indicate meta-information, provided the dataset is relatively homogeneous with respect to that information. In the new format of the IMIS database that documents the datasets in EMODnet Biology and EurOBIS, a table has been added that is capable of characterizing what functional group or what species list was targeted by each of the data sets.

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

An extract from the table characterizing the datasets containing benthos in the Southern North Sea (in total 180 datasets) is shown in Appendix C. We added a number of columns to the list of datasets, specifying:

- whether it is a collection of loose records or a dataset based on samples with several species
- What the main EOV targeted by the dataset is. Note that many plankton and fish datasets were included in the gross list, as they also report benthic species (e.g. as larvae for plankton, or as bycatch of epibenthic species in fish datasets)
- What the secondary EOV in the dataset is. This could be completed in such a way that all EOVs represented in the dataset are specified
- What size class was targeted. Only filled in for benthos (macro, meio, micro)
- What vertical division was targeted. Only filled in for benthos (hyper, epi, endo)
- If the dataset had sampled the complete target community

In case of doubt, 'completeness' of a dataset was tested by plotting the relation between effort (no. of sampling events) and number of species retrieved across all candidate datasets. We found that the effort-species points of the complete datasets all fell in a predictable linear range on a log-log plot, but that datasets known to sample only part of the community fell far outside this relation. We could restrict detailed checking of taxonomic coverage of datasets to those datasets with an aberrant effort-species value.

For incomplete datasets, we recorded the target list as the list of species represented in the dataset. It will have to be decided how to efficiently store this information with the datasets, but in principle this is possible.

Expanding on this experience, we propose to continue this classification of datasets as part of the work of WP4, but in close co-operation with WP2. The total number of datasets in EMODnet Biology is around 1000. This represents a considerable, but not an impossible amount of work.

With all datasets classified, this information can be fed back into the update of WoRMS as specified before. Presumably a number of iterations will be needed to complete the information, as scrutiny of more datasets will reveal new categories or new variables to be evaluated in the others. Where uncertainties remain, expert judgment or consultation of the data providers may be needed. Expert judgment on species can be obtained from the WoRMS taxonomic editors. Where the status of data sets remains unclear, a conservative attitude is to classify the dataset as a 'collection', so that it is only used to identify positive occurrences.

5 Application in products (Atlas of Marine Life)

We propose to prepare gridded and interpolated maps of species presence (and additionally density, biomass if possible) for a large number of species in different regional seas. This will rely on the completion of the metadata extension. It will first be applied to groups for which we have a reasonably good coverage and a good knowledge of the species within the WP4 group, and later be extended to other groups and species. Maps will be spatially restricted to areas that are well-covered with datasets and may represent different time periods when long time series are available. Error maps reflecting the confidence in the gridded fields will be produced. Such error

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

maps essentially depend on the data coverage, but also on the error on the observations, and can be used to blank our areas with too little confidence. Preparation of the data sets and gridding will be automated in a number of scripts that will be made available via GitHub.

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

6 Appendix A. Functional groups presently being added as characteristic traits to species

Benthos

By size

- **Macrobenthos** comprises the larger, more visible, benthic organisms that are greater than 1 mm in size. Some examples are polychaete worms, bivalves, echinoderms, sea anemones, corals, sponges, sea squirts, turbellarians and larger crustaceans such as crabs, lobsters and cumaceans.

- **Meiobenthos** comprises tiny benthic organisms that are less than 1 mm but greater than 0.1 mm in size. Some examples are nematodes, foraminifera, tardigrades, gastrotrichs and smaller crustaceans such as copepods and ostracods.

- **Microbenthos** comprises microscopic benthic organisms that are less than 0.1 mm in size. Some examples are bacteria, diatoms, ciliates, amoebae, flagellates.

By type

- **Zoobenthos** comprises the animals belonging to the benthos.

- **Phytobenthos** comprises the plants belonging to the benthos, mainly benthic diatoms and macroalgae (seaweed).

By location

- **Endobenthos** lives buried, or burrowing in the sediment, often in the oxygenated top layer, e.g., a sea pen or a sand dollar.

- **Epibenthos** lives on top of the sediments, e.g., like a sea cucumber or a sea snail crawling about.

- **Hyperbenthos** lives just above the sediment, e.g., a mysid shrimp.

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

Plankton

By trophic group

- **Phytoplankton** (from Greek phyton, or plant), autotrophic prokaryotic or eukaryotic algae that live near the water surface where there is sufficient light to support photosynthesis. Among the more important groups are the diatoms, cyanobacteria, dinoflagellates and coccolithophores.

- **[Mixoplankton** possess the ability to feed in an autotrophic or heterotrophic mode, i.e. they can perform photosynthesis but also prey on other planktonic organisms. Within the mixoplankton some taxa can produce their own photosynthetic apparatus, whereas others 'steal' it from their phytoplankton prey. The mixotrophic state of most species is still obscure, although recent observations suggest it is quite widely spread in marine plankton. This group is not yet recognized in the current state of the database, but is an obvious extension in a next phase]

- **Zooplankton** (from Greek zoon, or animal), small protozoans or metazoans (e.g. crustaceans and other animals) that feed on other plankton. Meroplankton (eggs and larvae of larger nektonic or benthic animals, such as fish, crustaceans, and annelids) are included here, but can be distinguished based on further classification.

- **Bacterioplankton**, bacteria and archaea, which play an important role in remineralising organic material down the water column. Note that prokaryotic phytoplankton are also, based on taxonomic considerations, bacterioplankton. However, from a functional point of view it seems preferable to classify them with phytoplankton, and reserve the class 'Bacterioplankton' to heterotrophic or chemo-autotrophic groups.

- **Mycoplankton**, fungi and fungus-like organisms, which, like bacterioplankton, are also significant in remineralisation and nutrient cycling.

By size

- **Megaplankton**: > 20 cm; metazoans; e.g. jellyfish; ctenophores; salps and pyrosomes (pelagic Tunicata); Cephalopoda; Amphipoda

- **Macroplankton**: 2-20 cm; metazoans; e.g. Pteropoda; Chaetognatha; Euphausiacea (krill); Medusae; ctenophores; salps, doliolids and pyrosomes (pelagic Tunicata); Cephalopoda; Janthinidae (one family of gastropods); Amphipoda

- **Mesoplankton**: 0.2-20 mm; metazoans; e.g. copepods; Medusae; Cladocera; Ostracoda; Chaetognatha; Pteropoda; Tunicata

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

- **Microplankton**: 20-200 µm; large eukaryotic protists; "net phytoplankton"; Foraminifera; tintinnids; other ciliates; Rotifera; juvenile metazoans - Crustacea (copepod nauplii)

- **Nanoplankton**: 2-20 µm; small eukaryotic protists; Small Diatoms; Small Flagellates; Pyrrophyta; Chrysophyta; Chlorophyta; Xanthophyta

- **Picoplankton**: 0.2-2 µm; small eukaryotic protists; bacteria; Chrysophyta

- **Femtoplankton**: < 0.2 µm; marine viruses

Other plankton

- **Meroplankton versus holoplankton.** Meroplankton is composed of eggs and larvae of nekton or benthos – species that do not spend their entire life cycle in the plankton. Holoplankton, in contrast, does spend its entire life cycle in the plankton
- **Ichthyoplankton.** Eggs and larvae of fish. This is a sub-group of meroplankton

Nekton

Nekton or necton refers to the aggregate of actively swimming aquatic organisms in a body of water.

Pleuston

Pleuston are the organisms that live in the thin surface layer existing at the air-water interface of a body of water as their habitat. Examples include some cyanobacteria, some gastropods, the ferns Azolla and Salvinia and the seed plants Lemna, Wolffia, Pistia, Eichhornia crassipes and Hydrocharis. Some fungi and fungi-like protists may also be found.

[Neuston

The term **Neuston** is used in ambiguous ways in the literature and is not used in the classification**]**

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

7 Appendix B. Current work to expand WoRMS with information gathered from data mining

7.1 Work done by checking the metadata of the datasets

We assigned functional groups (e.g. benthos, macrobenthos, plankton, phytoplankton, epibenthos) to EurOBIS datasets, based on the information in the title, abstract, description, data themes and keywords.

- E.g. if the title contains "macrobenthos", then it was assumed all species in this dataset can be considered as macrobenthos, and the functional group macrobenthos was assigned to this dataset.
- If the title says "benthos", but then the description mentions "macrofauna", then the functional groups "macrobenthos" and "zoobenthos" were assigned to this dataset.

It is possible to assign several functional groups to one dataset. Note, however, that not all datasets could get assigned a functional group. The resulting table with dataset ids (dasids) and functional groups was added to the EurOBIS database. The functional groups were also added as possible 'dataset theme' values in the metadata database 'imis', but not yet linked to the datasets. This can however be done quite quickly.

We then selected all datasets with only one or non-conflicting functional groups (e.g. zoobenthos and macrobenthos) and extracted the number of all species in those datasets (AphiaID's). These species lists are used in the final step to compare EurOBIS and Aphia information.

7.2 Work done by checking the Aphia database (WoRMS)

We checked what functional group information is already available in the Aphia database, the database behind the World Register of Marine Species (WoRMS), with a focus on the Belgian and European taxa:

- For all BeRMS taxa (Belgian Register of Marine Species), it was checked if the existing traits in Aphia could be used to refine the functional groups: e.g. if taxon had "benthos" and "5 mm length" as trait information, this was combined into "macrobenthos". Life stage and gender were taken into account for this. For taxa with body size information, the body size class was also assigned: Microbiota (<0,2 mm), Meiobiota (0,2 - 2,0 mm), Macrobiota (2,0 -200 mm) and Megabiota (> 200 mm). These size classes will be added as a new trait in Aphia as well.
- For all ERMS (European register) species, the body size class was assigned to each species (taking into account life stage and gender): Microbiota (<0,2 mm), Meiobiota (0,2 - 2,0 mm), Macrobiota (2,0 -200 mm) and Megabiota (> 200 mm).

7.3 Combining the EurOBIS and Aphia information

The results from EurOBIS and Aphia were then combined to find out if we could find functional group information in EurOBIS that is not available in Aphia yet. For this comparison we focused on accepted marine species. This exercise indeed resulted in a list of several thousands species for which we could add additional functional group information to Aphia. This information will

EASME/EMFF/2016/0066 – EMODnet Biology Phase III
D4.5: Proposed list of data packages and underlying datasets
with specific recommendations on (meta)data gaps

now be forwarded to the taxonomic editors, to ask if this can be added to the WoRMS (Aphia) database.

8 Appendix C. Some example dataset properties used in this report

dataset_id	collection	EOV.MAIN	EOV.SECOMD	EOV.SIZE	EOV.Habitat	COMPLETE	title
005577bc-da03-42c0-99d1-97b5e48cd06a	FALSE	PLANKTON	BENTHOS			TRUE	Semi-quantitive microplankton analysis (Sylt Roads Time Series) in the Wadden Sea off List, Sylt, North Sea in 1998
00a7f490-7418-413b-a9f1-666f186769b6	TRUE					NA	Museum and Art Gallery of the Northern Territory Malacology Collection - marine records
00f47c38-6233-4d6c-bd93-67a771ab6224	TRUE					NA	Collection Polychaeta SMF
03d594dd-3665-49e5-bb4c-a78b1528a8e3	FALSE	BENTHOS		MEIO	ENDO	TRUE	Study of the meiobenthos from a dumping site in the Southern Bight of the North Sea
057859e2-3ff2-4f55-950f-62089486a7f7	FALSE	BENTHOS		MACRO	ENDO	TRUE	Macrobenthos data from the Doggerbank - 2000
0b6c1359-22b7-46b6-bfb0-b67afaf7f245	FALSE	PLANKTON	BENTHOS			TRUE	Continuous Plankton Recorder (Phytoplankton)
0bae1308-d1f9-4c88-926b-50a16f1b7bde	FALSE	BENTHOS		MEIO	ENDO	TRUE	Cefas07 - Effects of simulated deposition of dredged material on structure of nematode assemblages - the role of burial
0c1cb7e9-c7d1-4643-b245-daa8839a183f	FALSE	BENTHOS		MACRO	ENDO	TRUE	The UK Archive for Marine Species and Habitats Data
0f7b0c9c-d77c-4852-9317-6a70c43a8031	FALSE	PLANKTON	BENTHOS			TRUE	Semi-quantitive microplankton analysis (Sylt Roads Time Series) in the Wadden Sea off List, Sylt, North Sea in 1999
10fe54bb-b93b-4f00-bf4c-bd51542614ef	TRUE					NA	Taxonomic Information System for the Belgian coastal area
119099f9-1de1-4f2a-9dc2-d67cb50c521d	FALSE	PLANKTON	BENTHOS			FALSE	Report on the Danish Oceanographical expeditions 1908-1910 to the Mediterranean and adjacent seas - PelagicPolychaetes
12381619-5331-4665-8abf-df4bf4763e9a	FALSE	BENTHOS		MACRO	ENDO	FALSE	Dutch national shellfish monitoring in the coastal zone
123d29bb-977b-4ea3-9c58-46cf79ed1523	FALSE	BENTHOS		MEIO	ENDO	TRUE	TROPHOS/PODO-I work-database I (23/01/2004): Meiobenthos from station 330 - structural and funtional biodiversity on the Belgian Continental Shelf
13abd5c3-ed8c-494c-9121-486484c8b63c	TRUE					NA	TWorsfold CullercoatsBay 2003
17cb6ba1-1b99-4715-9d8f-1a435537928a	FALSE	BENTHOS		MACRO	EPI	TRUE	Spatial and temporal epibenthos and hyperbenthos

							variations at the Belgian Continental Shelf monitoring stations
1d198847-58cf-4350-9d2c-8c8f611083c4	TRUE					NA	BEWREMABI dataset: Belgian Shipwreck - hotspots for Marine Biodiversity: Macrofauna on shipwrecks