EMODnet

European Marine Observation and Data Network

# EMODnet Biology

**EASME Identifier**

**Start date of the project: 19/04/2017**

## EMODnet Phase III

## D3.1: Improvements of the data entry procedure for data archeology and rescue

**Reporting period *(if applicable)*: 19/04/2017 – 19/04/2018**

*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 [ref]* | D3.1: Improvements of the data entry procedure for data archeology and rescue |
| WP title [ref]* | WP3: Data archaeology and rescue |
| Task [ref]* | |
| Authors [affiliation] | Nicolas Bailly, Dimitra Mavraki, Stamatina Nikolopoulou |
| Dissemination level | Public |
| Submission date | 16/03/2018 |
| Deliverable due date | 19/07/2017 |

*[ref] refers to the corresponding abbreviated name of the Deliverable (or WP, or Task...), if appropriate

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

# Contents

# 1 Overview

During the previous phase of EMODNET (EMN2), data were entered directly in the Darwin Core format (DwC). Data managers were trained through workshops and teleconferences.

Two documents were produced during EMN2 related to this activity (some complements to these two documents were inserted herein):

- A technical document with indications on metadata creation, and quality control (Mavraki, 2014).
- An article that was based on the report of a workshop conducted jointly between EMODNET and LifeWatchGreece, held in the Hellenic Centre for Marine Research in 2014 (Faulwetter et al., 2016).

Two main difficulties were recognized with the use (final report of EMN2 WP4):

- During the data entry: data providers have difficulty to stick to the standards when, 1) analyzing the data in the original document to fit them in the DwC standard, and 2) managing data that are not in the typical original format of the document (which are frequent in the expedition reports).
- During the data control by MedOBIS: the DwC is a flat format that lead to a great amount of repetition of data. The MedOBIS last step of quality control had necessarily to check all this repeated fields through all records to ensure that the values were the actually same, adding significant quality control working time.

The present document proposes one solution to address the two issues. Some elements of the two documents cited above were integrated to facilitate the reading.

It is based on two recommendations that were made in Mavraki (2014), which are followed more deeply than during EMN2:

- Never delete any data! Keep separate copy of working datasets!
- Always document your changes! Never assume anything!

The key ideas are:

- to enter data in an electronic format that mimics the layout of the document as much as it is possible: it facilitates data cleaning and data entry quality control (comparison of the data as in the original document and the digitized data).
- to customize data entry templates through a cooperative work between the data provider and MedOBIS.
- to transform/copy the digitized data in a pseudo-relational data schema represented by the customized templates.
- to match the customized templates with the DwC through a cooperative work between the data provider and MedOBIS.
- to leave the final production of the DwC file under the MedOBIS IPT from a clean, standardized, and quality-controlled dataset (whatever the starting format is).

This report is mainly the result of the EMN2 experience, and the procedure is intended to be tested with the dataset that received a grant in EMN3.

Reminder: Datasets that are targeted here are those built before 2000, especially before 1980 when the usage of (personal) computers was not spread in biodiversity.

# 2 Search and identification of potential datasets

Potential datasets may be under various formats from non-structured texts to structured tables.

The search starts by:

- Bibliographic search
- Search technical documents in research bodies related to marine biology. Some of those may have catalogues.

First step, documents are first sorted on their title. Second step, documents are checked one by one to verify that actual raw data are available in the document itself.

Data maybe under various data format:

- Disseminated in the text, the worst case, e.g., occurrence data mentioned along the text or in footnotes.
- In semi-standardized text: the data are displayed with the same layout, e.g., species accounts with the list of occurrences
- Presented in tabular layout but columns or rows can be subdivided irregularly.
- Presented in regular tabular layout.

Data may be available in electronic format as images, texts, spreadsheets, or exceptionally, databases (for old datasets).

# 3 Create the metadata

## 3.1 Ownership

Metadata has several technical purposes: (a) to discover the dataset(s) of interest; (b) to evaluate them for suitability for the intended purpose; (c) to understand them in context (i.e. the dataset related to projects, persons, organizations, facilities equipment, publications…); (d) to allow interoperation in the sense of a homogenous query over heterogeneous datasets (Mavraki, 2014).

They also are important to state the ownership as authorship and publisher of data and to indicate the potential contact persons that can help in the management of the dataset. In the case of old datasets, it might be difficult to establish the ownership, however, all efforts should be made to establish it.

Research institutions have not often maintained a catalogue of datasets which are published under their authority, especially for the past ones. Finding a service or a person in the institution that can give information is often difficult. It is reminded that ownership is independent of copyrights: data cannot be copyrighted but datasets may be. Each case must be treated carefully.

## 3.2 Step by step manual

Mavraki (2014) gives a detailed account for metadata entry.

It is an important step of dataset processing. During EMN2, many exchanges were necessary with the data managers to establish them properly.

One case during EMN2 was difficult. One dataset was created for several decades of sampling, which was performed with various gears, at different depth series, and targeting different species. The collation of metadata and the execution of quality control were subsequently very difficult to perform. The full dataset was eventually split in several subdatasets to facilitate the integration in MedOBIS. Mavraki (2014) indicates in these cases: it is recommended that the description indicates whether the dataset is a subject of a larger dataset and provide the related link to the parent metadata/ dataset

# 4 Adapt the data entry schema to the format of the original dataset vs DwC

At first glance, this phase may seem to lengthen the working time of the total data entry process, from the original document to the final database and its exports in DwC format. However, a lesser quality wok at the start generates more work during the final quality control performed the OBIS node.

There are two steps:

- Digitize the data in a format that mimics the layout of the original document: 2 first subsections.
- Preparation of the templates for data entry: third subsection.

The issue in terms in work management if that the data entry process does not produce record after record, but batches of record: it is difficult to evidence the work done by the number of records entered per day, which might be a problem in certain contexts. As for the whole data entry exercise, the work must be adapted to the current situation.

## 4.1 Data to extract from the literature

To improve this step, the recommendations of Mavraki (2014) were pushed further: "Never delete any data! Keep separate copy of working datasets!"; "Always document your changes! Never assume anything!".

The key idea is to enter data in an electronic format that mimics the layout of the document as much as it is possible: Two main advantages:

- When data are in electronic format, it is much easier to handle them by batch: sorting, find and replace, formatting, formulas, queries, etc. All functionalities usually available from word processors, spreadsheets, database management system are available to proceed step by step with the standardization procedure, e.g.: if measurement are either in mm or cm, search in the document allows to change to the standard unit all at once, and not one by one; completing genus name when they are abbreviated; correct the layout of geocoordinates; etc.
- The 'undo' functionality is available in all these applications (CTRL Z), which allows to roll back in case of doubt on the already performed operations without being to start all over. In addition, versions of the file should be saved regularly, especially before operations that would result in many changes.

If the document is first scanned with a character recognition (OCR), then this methodology is even more important, as the OCR results may result in several mis-recognitions: confusion between '1', 'l', 'I', between 'mr' and 'nn', between 'cl' and 'd', and others, the more the older the printed document is. As mentioned in Faulwetter et al. (2016), some old documents are not worth a scan-OCR step, especially when characters (letters and digits) are not well separated or the background is too dark.

### *4.1.1 From text*

For data spread all over the text, there are few chances that this step is useful.

In some cases, a format can be pre-established and filed with data as the document is read. But not too much time should be spent for this because usually, as variations and exceptions to a pre-established format are encountered along the data entry work progresses.

For data that are semi-structured in the text, the use of the regular expression that allow to match some textual patterns helping in moving in the same text file for a semi-structured to structured data format.

### *4.1.2 From tables*

Reproducing the format of the printed document has a great advantage. It allows to check visually the result of the data entry in the strictest way, with a significant gain of quality and time.

## 4.2 If data are already under an electronic format

If the document is already in an electronic format, it does not mean the data can be extracted automatically. It depends on the underlying file format.

For instance, PDF file may present different formatting. Aside form the case of images, a two columns text maybe be formatted by column and the extraction is direct, or as lines extending to the two columns. The is a step of recovering the layout format that may be not negligible. In other cases, some texts in PDF, are the results of an OCR phase (see above), the text that can be extracted not corresponding to the text that is displayed. These last two are to be taken care of for older documents.

## 4.3 Preparation of the templates for data entry

The key point of the revised procedure is that an important collaborative work must be carried out jointly by the data provider and MedOBIS (taken as an example of any OBIS node).

Several milestones in that step may be followed for the treatment of the metadata, and the preparation of the data entry schema, depending on the skills of the data provider, here the data provider is assumed to be a beginner (hereafter, validated means a final quality control is performed by MedOBIS):

- The original document is shared to be studied by both the data provider and MedOBIS.
- First teleconference where MedOBIS explains the whereabouts and the technicalities of the data entry. MedOBIS provides some templates that seems to fit best with the structure of the document.
- The data provider creates a template for data entries from scratch or modifies or adopts one provided template.
- Second teleconference where the templates are discussed and tested. Potential deviations to templates are examined and decision are taken on how to deal with such cases. Each of the following steps may require several email exchanges and dedicated teleconference on specific issues.
- The metadata are delivered and validated by MedOBIS. The information that must be attached to the occurrence record are organized in a table (particularly, these are the data that are repeated in the DwC during EMN2: they are here entered and validated one only).
- Data entry *per se* (see next section).

This procedure has to be adapted case by case depending on the structure of the original document where the data are extracted from, and the skill of the data provider. Some data providers may have their own database system, and enter the data first in their schema, the work with MedOBIS in then shifted in the section 'Data extraction from structured file'.

# 5 Data entry

As mentioned in Faulwetter et al. (2016), one preliminary recommendation is to read carefully the introduction in the document, explaining the whereabouts of the occurrences datasets, the followed standards, the exceptions to a general schema, any information that can be used in the metadata, or some remark fields, etc. For the results of expeditions which are published in several volumes / issues, read carefully the volume where a narrative of the expedition is given and the stations. The stations maybe in a separated issue as well. During the data entry, the data manager should have these documents at hand.

This step corresponds to the data entry / transfer to a pre – relational model format. The idea here is two enter the topic-related information together, not directly the complete record.

Typically for occurrence datasets, it is often easier to create several tables, one for each topic below:

- Taxonomy: includes para-taxonomic names such as common names.
- Location: sampling stations and localities of the observation or the collection event.
- Collection / sampling procedure
- Occurrences: Taxonomy (link or reference menu) X Stations (link or reference menu) + measurements (data entry) + specimen traits (data entry)

Several milestones in that step may be followed, depending on the skills of the data provider, here the data provider is assumed to be a beginner (hereafter, validated means a final quality control is performed by MedOBIS):

- The taxonomy list is delivered and validated by MedOBIS. WoRMS is alerted for missing names.
- The location / station list is delivered and validated by MedOBIS.
- If necessary, a dedicated step on environmental parameters and specimen traits is added.
- Third teleconference (2 first in the precedent section) to finalize the template for occurrence records entry: creation of drop down menus from the taxonomy and locations / stations lists.
- The full dataset is provided after an overall quality control by the data provider.
- Overall quality control by MedOBIS.

This is the general principle, and again, one must adapt to the situation case by case.

## 5.1 Recommendations on data entry

### *5.1.1 Taxonomy*

It is recommended to prepare a taxonomic reference list that fits the document.

It can be prepared from WoRMS, e.g., if an expedition is reported from Greece, prepare the list from the species recorded in Greece and complete with names that are not in WoRMS (signal those to WoRMS immediately so that AphiaIDs can be created before the finalization of the dataset entry).

In many datasets, there are collective names, common names, etc. that do not match with any taxon in WoRMS. They must be added in the taxonomic reference list to ease the occurrence data entry. It explains why WoRMS cannot be used straightforwardly, in a two columns table: one with the name as used in the document, one with the corresponding name in WoRMS and/or AphiaID.

### *5.1.2 Locations*

Generally, the locations correspond to the sampling stations. Preparing a list of locations helps to clean and standardize the geocoordinates, and to complete the information required by the DwC (country, ocean, etc.).

### *5.1.3 Collection / sampling procedures*

The list of collection and sampling gears and procedures should be built at the same time as for the metadata. Some discrepancies may occur between an introductory or material and method section, and the data indicated in the occurrences themselves, they must be resolved at that step. Metadata have to be corrected accordingly.

### *5.1.4 Environmental parameters*

Likewise, a list of environmental parameters and methodologies should be constituted along with the metadata. Units are to be particularly taken care of. It seems to be an obvious recommendation, but in the daily work of data entry that may become a less critical repetitive operation, some errors happen. Hopefully, further quality controls will reveal these mistakes.

### *5.1.5 Species / Specimen traits*

Species traits must be differentiated from specimen traits. Usually, and unless a specific occurrence dataset has been created in purpose, information about the collected specimens / observed individuals are entered as remarks.

If those are important in size, the data manger has to decide if these information are worth to enter related to the goal of the dataset, or ignore them, at least in a first phase, privileging speed over completeness to deliver the dataset. Note that any indication helping to confirm the identification should be entered no matter the goal of the dataset.

## 5.2 Technical Quality control during the data entry

Two strategies can be adopted for the data control.

- A final one after all data are entered. This avoids repeating the same procedures several times.
- Perform a technical quality control regularly. Even if the first strategy, it happens irregularly as the data manager realizes that something looks wrong. But here its is to organize some quality controls during the data entry.

The second strategy may seem to take more time, but for huge datasets, it is preferable. It sometimes helps to adapt slightly the templates on the fly to make the data entry easier.

As usual, strategies may be adapted case by case.

## 5.3 Detecting and correcting errors in original data

When entering data, some errors in the original document are detected. The recommendations of Mavraki (2014) is to keep the data verbatim and never assumed that it is a mistake.

However. Depending on the skills of the data provider, he may want to correct those.

The recommendation is to keep a verbatim version of the files, and to work on a corrected version, e.g.:

- If the original file is a text, highlighting the corrections or annotated them in the corrected version.
- If the original file is a spreadsheet or a database, add a column with the corrected data, and continue the procedure with these columns. It is a good practice to create one anyway where the starting, trailing and double spaces are removed. And flag the records corrected. In particular, not all scientific name misspellings are to be entered in WoRMS: they are corrected, and the original spelling put in remark.

In a final step, these corrections may be reported in the corresponding record if relevant to the occurrence.

## 5.4 Detecting and correcting data encoding errors

Ultimately, they should be detected in the successive quality procedures. These ones are better detected if a slight quality control is operated daily, or on subdatasets when identifiable.

Another method that may be use: reconstruction the data under the same layout than the original document, and comparison the two electronical documents. It is not always possible, and may take some time, all depends on the structure of the original data and the importance to have data as clean as possible.

Double entry is another method used for strategic datasets in other domain. It seems unrealistic in biodiversity domain.

## 5.5 Three possible steps loop

Depending on the format of the original document, it may be interesting to use a suite of applications with different functionalities such as:

- Word processor: using the regular expression to match text pattern allows to transform a text in lines to structure columns in certain cases, to standardize some names, …
- Spreadsheet: the functions allows to merge or split columns according to DwC field standards.
- Database: using queries allow to perform regular (daily) technical quality control

This loop mainly allows entering correcting data at one place only, and to make copy only electronically, compared to cases where the same information is entered several times, minimizing the data encoding errors (typos); it strengthen the quality level.

If the files mimicking the original document are important for any reason (e.g., dissemination), mistakes detected from queries in a database can be corrected in the word document, which restart a cycle.

This requires that the data provider is skilled in these different applications, knows how to transfer files from one to another (there are less and less problems with the management of character coding but there are still some surprises), and the original structure does not differ too much from the final one (unless data quality is privileged over the data entry speed).

## 5.6 Test of conversion to DwC

In the case where the conversion to DwC is already available, regularly converting the data entry dataset allows to check visually the output and correct obvious errors, before the final quality control.

## 5.7 Recommendation for correcting errors at each step

The final recommendation is to make all efforts to correct all issues at the end of each step.

This avoids a final quality control that may be very complicated and time-costly in the end.

# 6 Data extraction from a structured file (spreadsheet, database, csv)

When the dataset is already under a structured format, several strategies can be adopted depending if the structure match more or less with the DwC, and if all or part of the data are to be extracted.

Two main opposite ones are listed here.

- Matching the structure directly to the DwC, especially if the dataset is well curated and quality-controlled. From a MedOBIS side, it requires some work to integrate the data in the MedOBIS database.
- Starting the procedure as described here, by creating the taxonomy, locations / stations, etc. reference tables: it may be the preferred choice if only a part of the data are to be extracted or if the structure is not close to the DwC.

# 7 Integration of data in the OBIS node database (here MedOBIS)

The key point here is related to the fact that the data provider may not be familiar with the DarwinCore. Training is worth only if the data provider will use it regularly. In any case, the MedOBIS is better knowledgeable than the data provider and will know better how to fit the dataset in the DwC standard. Hence the time gained (but apparently wasted) at the start of the procedure in the cooperative work between the data provider and MedOBIS.

As much as possible, the customized templates should be elaborated as close as possible of the structure of the MedOBIS database that will host the final version of the dataset. Then the transformation in DwC is realized as a general routine for the IPT.

# 8 Semantic Quality control in the OBIS node database

The OBIS nodes are in charge of the semantic quality control. While it is conducted along the procedure during various steps, a final quality control is performed.

## 8.1 Assignment of AphiaIDs

In particular, the entry of new names in WoRMS may require some time, depending on the reactivity of the relevant taxonomic editor. MedOBIS will conduct regular checks for names that were still missing at the time of the integration. Meanwhile, they will be replaced by the lowest rank parent taxon that existed in WoRMS at that time.

## 8.2 Using the LifeWatch Belgium QC tool online

Depending on the skills of the data provider, the LifeWatch Belgium Quality Control tool will be introduced to the data provider so that the tool can be used along the steps of the procedure.

# 9 References

FAULWETTER, S.; PAFILIS, E.; FANINI, L.; BAILLY, N.; AGOSTI, D.; ARVANITIDIS, C.; BOICENCO, L.; CAPATANO, T.; CLAUS, S.; DEKEYZER, S.; GEORGIEV, T.; LEGAKI, A.; MAVRAKI, D.; OULAS, A.; PAPASTEFANOU, G.; PENEV, L.; SAUTTER, G.; SCHIGEL, D.; SENDEROV, V.; TEACA, A.; TSOMPANOU, M.; 2016. EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases. Research Ideas and Outcomes, 2(e9774): 1-28. doi:10.3897/rio.2.e9774

Mavraki, Dimitra; 2014. Marine Data Quality Control Manual. EMODNET2 Technical document. Heraklion (HCMR): HCMR. 10 p.