EMODnet

European Marine Observation and Data Network

EMODnet Biology

EASME/EMFF/2016/006

EMODnet Phase III

D3.4: Policy report on biodiversity data management sent to research organizations

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	D3.4: Policy report on biodiversity data management sent to research organizations
WP title	WP3: Data archaeology and rescue
Task	Task 1: a common method of access to data held in repositories
Authors	Nicolas Bailly (HCMR) <nbailly@hcmr.gr>
Dissemination level	Public

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

Contents

D3.4: Policy report on biodiversity data management sent to research organizations

Data are the basic 'material' of science as phenomenon measurements resulting from observations in the natural world and laboratory experiments. Yet data management seems not to be integrated enough in biodiversity and environmental sciences. Building an e-infrastructure such as EMODnet is hindered by the low quality of datasets that are provided: too many small deviations from the followed standards, approximate quality control, lack of detailed metadata, etc. Correcting these datasets are costly in time while minimum efforts by data owners from the start would considerably speed up the integration process in big data repositories.

The principal reason is that biological scientists are still not trained enough to databases procedures and related technologies, and consequently data management is neglected not only during projects but also especially after their end, whatever researchers have to do it by themselves for personal works, or to hire professional data managers for larger projects.

A number of European directives such as INSPIRE already exist, but their implementation on the ground are lagging behind. Strong political commitments from EC, countries and their research institutions are required to raise up the data management profile at all levels: student education, scientist training, research team and institution organization related to data management. This document is also an urgent recommendation to research institutions and research managers to develop data management policies and procedures on the ground, and to hire professional staff required to implement them in order to facilitate and improve Big Data take up.

Before computers, data were reported on paper publications usually under the form of tables with columns and rows that were explained in the text of the article, which scientists were trained to. When computers and then databases emerged however, scientists especially biologists were still trained to report their data on papers, but not to organize them in documented spreadsheets, and even less in databases. For example, publications describing datasets are not physically linked to the spreadsheets and data may be separated from their explanations: methods, units, context of experiments or observations, all information that are now termed metadata (data about data).

Databases, and more generally datasets, are key components of the current Big Data trend in science, and of informatics infrastructures (e-infrastructures) that are currently developed: data repositories portals, analytical software and workflow platforms, virtual laboratories, virtual research environments, etc.

EMODnet, an EC DG-MARE project, is developing such an e-infrastructure dedicated to the marine environment in Europe, gathering data and analytical tools in a unique portal. In its module Biology, an important activity consisted in gathering and digitizing datasets about species occurrences that would otherwise remain difficult to access (e.g., grey literature) or be possibly lost (e.g., spreadsheet files). The digitization process forces better structuration and standardization of data to fit in a database from where they can be used altogether in large scientific analyses.

During EMODnet, historical data extracted from the literature were digitized, and more recent datasets were integrated from electronic files under various formats and structures. As one could expect, the

historical datasets were not simple to digitize, especially when they were not presented as tables. Paper reporting allows many deviations from a standard, and various processes to reduce the space on paper were used. They must be taken care of during the digitization phase, which takes time. Moreover, when explanations are too succinct, the quality and usability of data are definitely compromised.

Surprisingly, the situation was not much better for recent datasets in many cases. Progressively it was realized that despite the existence of personal computers for 40 years, biologists are not trained and organized to build electronic datasets / databases than can be easily re-used by others. Luckily, data owners could be contacted in order to clarify issues, but it is usually a long process that slows down the delivery of the final dataset under a proper format for integration in the EMODnet repository.

Spreadsheets are the easiest way to electronically record data when one has little knowledge of informatics: it looks like a table on paper, easy to handle. For technical reasons beyond the scope of this document, true electronic databases should be preferred, but they are more difficult to understand. The relational model that underlies most of the database systems today is not intuitive and present some caveats requiring a thorough training. That is the reason why there are usually user interfaces that mask some of the difficult deep technical parts.

Scientific students must be educated during their undergraduate and graduate studies. Through exercises, they should be able to build simple databases for educative projects, and to report properly the associated metadata. They should also be confronted to exploiting databases of uncertain quality and metadata for analyses in order to further understand the cost of not having properly quality-controlled and documented datasets.

Scientists in position must be trained though a number of workshops to realize the importance of delivering proper datasets immediately exploitable by students and colleagues, and Big Data repositories, beyond the publications. Data management should be part of any project proposal and evaluation.

Team leaders and unit directors must organize the data management at their level. Depending on the size of the units, they should ensure that professional data managers are hired full, or part-time thus being shared by several units. Another option is that the institutions organize the data management at that level, providing data manager time to units when needed, especially at the beginning to structure an adapted and efficient database, but also after the end of projects to properly curate the final dataset in terms of quality control and metadata writing.

Research institutions must implement policies and provide facilities for efficient data management. An important point is to integrate data management activities in the job descriptions, that will be part of the staff work evaluation, but also team and institution evaluations. In terms of project management, institutions should make sure that datasets are properly delivered at the end of projects, and provided to a data repository, either institutional, national, or international. And a catalogue of the datasets available from the institution should be put online.

Actions must be undertaken swiftly to facilitate the rapid further development of efficient e-infrastructures that will help to solve complex biodiversity and environmental issues that for marine environment are directly related to the Blue Economy and its Growth.