Marine biogeographic data in EurOBIS: assessing their quality, completeness and fitness for use

Leen Vandepitte, Flanders Marine Institute, leen.vandepitte@vliz.be (Belgium)
Filip Waumans, Flanders Marine Institute, filip.waumans@vliz.be
Lennert Tyberghein, Flanders Marine Institute, lennert.tyberghein@vliz.be
Bart Vanhoorne, Flanders Marine Institute, bart.vanhoorne@vliz.be
Francisco Hernandez, Flanders Marine Institute, francisco.hernandez@vliz.be

The European Ocean Biogeographic Information System - EurOBIS - is an online marine biogeographic database compiling data on all living marine creatures (www.eurobis.org). The principle aims of EurOBIS are to centralize the largely scattered biogeographic data on marine species collected by European institutions and to make these data freely available and easily accessible.

The available data are either collected within European marine waters or by European researchers and institutes outside Europe. The database focuses on taxonomy and distribution records in space and time; all data can be searched and visualized through a set of online mapping tools. All data are freely available online and easily accessible, without requiring a login or password.

Given the very diverse nature of the data - going from museum collection data to literature data and research and monitoring data -, the standardization of both the data and the data format and evaluating the quality is not always evident. To simplify this task, a set of quality control procedures have been developed, encompassing taxonomic, geographic, outlier and data format checks.

The aim of these quality control procedures is two-fold. First of all, it helps the data management team and data providers to check the quality and completeness of the submitted data and detect (possible) errors. Records not meeting the assumed quality standards are sent back to the provider for a secondary check-up, clarification and/or corrections. This back-and-forth communication between the provider and the data management can greatly improve the quality of the submitted dataset, thereby also enhancing the quality of the data available in the integrated data system. Secondly, the assigned quality flags can help users in selecting data from EurOBIS that are fit for their use and purpose.

On the data management level, each individual distribution record is submitted to 20 quality control (QC) steps, generating 20 quality control flags. These steps include e.g. a check on the completeness of the required fields of the OBIS data scheme, a verification of the filled in values for the date related fields (e.g. is the month value between 1 and 12 or does the start date precede the end date) or checking whether the given sampling depth is a possible value compared to existing depth profiles. One of the most important checks is related to the taxon name of a record:

any taxon name within EurOBIS should be matched to the World Register of Marine Species (WoRMS, www.marinespecies.org), the most authoritative and comprehensive list of names of marine organisms, including information on synonymy. Only be linking to WoRMS, it is possible to rule out spelling variations and synonyms. This allows grouping of distribution records in a reliable way for further analyses. If a taxon name does not appear in WoRMS, this is further investigated: If it would consider a marine taxon not yet present in WoRMS, the appropriate taxonomic editors are contacted, so the taxon can be added and linked. If the name is from a non-marine taxon or does not make sense, it is added to an annotated list, explaining why it is not documented in WoRMS.

Next to these 20 QC steps on record level, additional geographic and taxonomic outlier analyses are run on respectively the dataset and the entire database, generating 2 more quality control flags. The geographic outlier analysis compares the observation points within a dataset and identifies possible outliers.

This kind of check can reveal possible errors in the latitude and longitude values, e.g. because of switched latitude and longitude values or a missing minus sign to indicate south or west. The taxonomic outlier analysis runs on the entire EurOBIS data system and will identify species that are documented outside their normal occurrence range, which is based on the available data within EurOBIS. These species outliers will need further investigation and verification, as they can be actual outliers or they can be new occurrences of the species in a previously undocumented geographic area. Both these outlier analyses will thus help in assessing the validity of a record compared to all available distribution records within one dataset and within the entire EurOBIS data system.

All these automated quality control steps will also be implemented on the international OBIS data system, greatly improving the 'fitness for use' of marine species distribution data on an international level.

The assigned quality control flags can be combined according to the required 'fitness for use' for the users, thereby creating specific filters on the available data within EurOBIS. The European Marine Data and Observation Network (EMODnet) Biology Portal (http://bio.emodnet.eu/portal) is currently applying such a filter. It only makes available those distribution records that comply with the following QC steps: the required fields - according to the OBIS data schema - are completed, the taxon name relates to a genus or species and is listed in the World Register of Marine Species (WoRMS), and the coordinates are – format wise – correct. On the dataset level, users can see how many records have passed the postulated quality control procedures and are thus available through the portal.

For the individual data providers, a number of the developed quality control procedures are offered as a web service through the LifeWatch Portal (www.lifewatch.be), enabling researchers to run certain checks on their data themselves. These web services currently encompass a data format check, a geographical check, a taxon name check and a check for (possible) duplicate records. The results of each check are directly available in an output file, so the provider can immediately check and – where necessary – adapt the uploaded file. The taxon check not only includes a match with the World Register of Marine Species (WoRMS), but also has look-up functionalities for other taxonomic databases such as e.g. the Integrated Taxonomic Information System (ITIS), the Catalogue of Life (CoL) and the Integrated Register of Marine and Non-Marine Genera (IRMNG).

Checking taxon names against several taxonomic registers will already help the provider in assessing the validity of the name and indicate possible errors (e.g. a terrestrial animal in a deep-sea dataset), before submitting it to EurOBIS.

The development of these QC procedures is part of the VLIZ contribution to LifeWatch, and funded by the Hercules Foundation. The main goal is to facilitate the fitness for use of individual and integrated biogeographic data for scientists, by offering several tools that help in the assessment of the completeness and validity of distribution records.