INTEGRATING ENVIRONMENTAL DATA FROM HETEROGENEOUS SOURCES:LESSONS LEARNED IN MARBEF

Edward Vanden Berghe
Flanders Marine Institute, Wandelaarkaai 7, 8400 Oostende, Belgium - wardvdb@vliz.be

Abstract
One of the objectives of the EU Network of Excellence 'Marine Biodiversity and Ecosystem Functioning' was to bring together existing datasets in large databases, to facilitate analysis on long term, and on the European scale. Data integration proved to be a non-trivial and labour-intensive task. Good standard lists (vocabularies/ontologies) for both taxonomy and geography are essential tools. Vastly different sampling equipment and design often precludes quantitative analysis on the level of the integrated database. Sampling bias, resulting from the objectives of the projects that generated the individual data sets, needs special attention.
Keywords : Analytical Methods, Biogeography, Ocean History.

Scientific data are often collected in the framework of relatively small projects; the resulting datasets are usually relatively small-scale, and fail to inform on the scale of global environmental problems with which humankind is confronted. This has prompted a multitude of data integration activities, to try and assemble several of these smaller datasets in larger, interpretable databases. One of the objectives of the EU Network of Excellence 'Marine Biodiversity and Ecosystem Functioning' was to bring together existing datasets in large databases, to facilitate analysis on long term, and on the European scale.

Older datasets are transcribed from paper to electronic forms in data archaeology and rescue projects. The digitisation of historical data from paper files can cost $\leq 0.5\%$ of that of the original field surveys [1]. While these integrating and rescue activities are extremely useful, they are often labour-intensive, they often cause a certain degree of loss of detail in the information, and they introduce an inherent danger of data duplication. Very often, essential metadata (such as sample size) are not immediately available with the dataset, and have to be traced in literature based on the data. Lack of standard protocols results in variations in sampling equipment, and makes quantitative comparisons difficult. Often sample size is missing, and abundances are only given as densities per unit area or volume - but this precludes the use of rarefaction, which is an essential technique when dealing with samples from different sizes.

To facilitate integration, reference should be made as much as possible to standard vocabularies of terminology: standard taxonomic names, geographical names and parameter names. For parameter descriptions, a dictionary developed at the British Oceanographic Data Centre is used [2].The reference for taxonomic names used within MarBEF is the European Register of Marine Species (ERMS) [3, 4, 5], a synonymised list of names of organisms that have been reported from European marine waters. ERMS contains not only valid names, but also invalid synonyms and documented misspellings; this way, ERMS forms a guide to the correct application of taxonomic names. ERMS provides an online tool for data integration. A gazetteer of marine place names was compiled, based on several existing lists of place names (e.g. from IHO and FAO), and on the basis of literature on an ad-hoc basis [4]. The purpose of the gazetteer is to improve access and clarity of the different geographic, mainly marine names such as seas, sandbanks, ridges, bays or even standard sampling stations used in marine research. Maritime boundaries and Exclusive Economic Zones in particular are important concepts for a lot of biogeographical applications. As no global public-domain cover of such information was available, the Flanders Marine Institute (VLIZ) decided to develop it: treaties between countries were gathered and the coordinates that were published herein were imported in a GIS. Where no treaties were available, maritime boundaries were calculated in ArcGis as 200 nautical mile buffer lines or as median lines, according to the regulations of the United Nations Convention on the Law of the Sea.

Where standard protocols were used in generating the observations or where there is an opportunity to make a priori agreements with the scientists about minimal required metadata, the integration process can be simplified and the loss of detail can be minimized. In all cases, extensive documentation is essential. Some controlled vocabularies exist, such as those provided by the Global Change Metadata Standard (GCMD), ISO 19115 and the Federal Geographic Data Committee (FGDC), but urgently need to be expanded for management of marine biology and ecology data.

Most data are collected using public funding of some kind; it is only reasonable to assume that these data would ultimately be available in the public domain. Unfortunately, that is not always the case; for example, NODCs contain less than half of the oceanographic data collected in their countries [6]. Few of the marine papers in top journals publish raw data. The concerns of data owners, and reasons why not to make data publicly available, were reviewed by Froese [7]. Incentives are needed for data custodians to share their data. One possible mechanism would be to make datasets citeable; contributing data to on-line datasets, and having the data cited by others, should be treated in the same way as publishing research papers and being cited. To facilitate this, authors of scientific papers could be requested to deposit their raw data in a public, well-managed archive. It is the expected practice in taxonomy to lodge type specimens in museums and, in genetics, to deposit sequences in GenBank, prior to publication. There should be a similar requirement by journals that ecological data be made publicly available prior to printed publication [8]. Apart from making data more widely available, this has the advantage of increasing the level of possible peer-review.

References

1 - Zeller D., Froese R. and Pauly D., 2005. On losing and recovering fisheries and marine science data. *Mar. Policy* 29: 69-73.
2 - Lowry R., Bird L. and Haaring P., 2004. A semantic modelling approach to biological parameter interoperability, In: Vanden Berghe E. (ed.), Ocean Biodiversity Informatics, Hamburg, Germany: 29 November to 1 December 2004: book of abstracts. pp. 14.
3 - Costello M.J., Emblow C. and White R., (ed.), 2001. European register of marine species: a check-list of the marine species in Europe and a bibliography of guides to their identification. Collection Patrimoines Naturels, 50. Muséum national d'Histoire naturelle: Paris, France. ISBN 2-85653-538-0. pp. 463.
4 - Vanden Berghe E., Bouchet P., Boxshall G., Costello M.J. and Emblow C., 2004. European Register for Marine Species version 2.0: data management, current status and plans for the future. In: Vanden Berghe E. (ed.), Ocean Biodiversity Informatics, Hamburg, Germany: 29 November to 1 December 2004: book of abstracts. pp. 29.
5 - Costello M.J., Bouchet P., Boxshall G., Emblow C. and Vanden Berghe, E., 2004. European Register of Marine Species. Available online at http://www.marbef.org/data/erms.php.
6 - Deckers P. and Vanden Berghe E., 2006. The VLIZ Maritime Boundaries Geodatabase as a biogeographical tool. In: Abstracts of the 2006 ICES Annual Science Conference 19-23 September; Maastricht, pp. 228.
7 - Kohnke D., Costello M.J., Crease J., Folack J., Martinez Guingla R. and Michida Y., 2005. Review of the International Oceanographic Data and Information Exchange (IODE). Report submitted to the Intergovernmental Oceanographic Commission (IOC) of UNESCO, 23rd session of the assembly. Available at http://ioc3.unesco.org/iode/files.php?action=viewfile& fid=501& fcat_id=124.
8 - Froese R., Lloris D. and Opitz S., 2004. The need to make scientific data publicly available - concerns and possible solutions. In: Palomares M.L.D., Samb B., Diouf T., Vakily J.M. and Pauly D. (eds), Fish biodiversity: local studies as basis for global inferences. Fisheries Research Report 14, ACP-EU, Brussels, pp 268-271.