CM2006/M:26

Revealing species communities in a spatial and temporal overlap

A. Meerhaeghe, K. De Cauwer, M. Devolder, S. Jans, S. Scory

Belgian Marine Data Centre, Management Unit of the North Sea Mathematical Models,
Gulledelle 100, B–1200 Brussels, Belgium
http://datacentre.mumm.ac.be/, e-mail: bmdc@mumm.ac.be

Abstract

Storing physical and chemical values, optical spectra and sediment granulometry is already a cumbersome task; dealing with biological data even falls into a different category. Biologists tend to focus their attention to species of their interest while other specimens in the same sample are often underestimated. Meanwhile taxonomists are continuously revising the taxonomy resulting in a complete new set of relations between these taxa. Keeping track with both and meanwhile having a dataset up to date seems endless. At the Belgian Marine Data Centre we tried to think outside the box and came up with a solution to content both biologists and data managers. The last thing we aimed at is to create another web index to refer species, therefore we hooked up with the existing web based referencing systems. The need to get data about different food webs in a spatial and temporal overlap is answered by our hierarchical storage of taxa which allows selecting a predator at species level and at the same time selecting different prey species at lower taxonomic levels. As these species, and also the scientists, usually are not confined into ‘latitude longitude squares’ we elaborated the spatial selection tool which defines user specific polygons to base the selection of data upon. We will briefly present the structure of our relational database but specific attention will go out to the taxonomic and spatial parts. Incentives and discomforts to organize the data in this way, and our current web interface, will be demonstrated.

Introduction

With a lack of space on the mainland for different human activities the demand for marine space increased the last years. As the economical and social pressure built up and ecological and international policy came into play as well, practices of terrestrial planning were adapted to the marine world. More in particular in the context of sustainable energy sources, far shore wind farms are a window of opportunity to plan the resources at the North Sea. In the zone under Belgian jurisdiction there is currently one company that got all the permits and approvals do build such a wind farm but others are coming. They plan to install 60 turbines (3.6-5 MW) on the Thornton bank, situated about 27 km from the Belgian coast. This permit includes a continuous monitoring programme to monitor possible effects on the marine environment beginning one year before the start of the construction

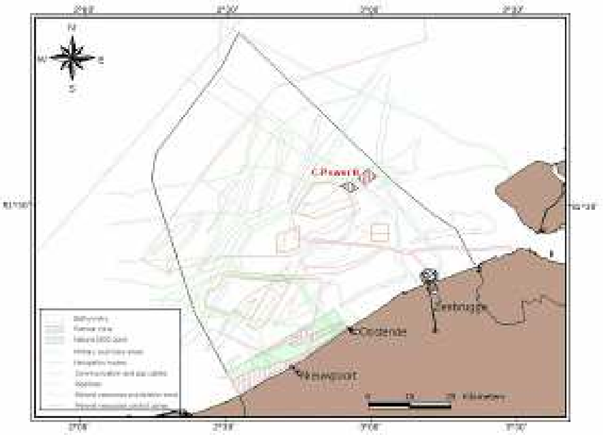


Fig. 1: Situation of the wind farm at the Belgian Continental Shelf in relation to other human activities

phase. With the expertise at the Belgian Marine Data Centre to manage marine data from different kind of scientific resources we were solicited to take the data management under our wings. In this paper it is not our aim to go into the detail of this project but we will focus on the biological part, *sensu lato*, of it and how we dealt with the initial problems after one year of monitoring. We will conclude with some future applications as possible spin-offs from this project, which could be reused in a much broader context.

Available research data

Biogeographical models of species' distributions and ecological modelling are gaining terrain when we build assessment studies about impacts of changing environmental conditions, linked directly or indirectly to human activities. Nevertheless we should keep in mind that the vast majority of these recent advances deal with specific terrestrial communities and species' assemblages. Extrapolation to other habitats on land (or at sea) or generalisation of these models should be considered with the greatest care possible. In aquatic ecosystems, and certainly in marine environments, limitations to dispersal are almost absent and the flux of energy and material within and between habitats tends to be faster in comparison with these on land. At the other hand the availability and the experience to manage big data sets in marine sciences is a major advantage to modelling; a trade off between these experiences 'at land' and 'at sea' has potential for novel insights in both domains (Raffaeli, 2006).

This principle of 'access to data' stands ground for physical, chemical and geological data of the marine environment. While these marine sciences evolved with the expanding possibility of the information age during the last years and they collect data in a semi-automatic way, marine biology still is 'fishing' with the same gear and data collection is, for its biggest part, manual. Biological data collected end usually up into one of the four following categories.

'Presence data' only provides information about the location of a given organism and has its origin often somewhere in a collection. It is common knowledge that natural history musea world-wide are harbouring an enormous treasure of information on past and present distribution of plants and animals. Unfortunately these 'records' often are neither digitised, nor geo-referenced. Once these records are made available they would provide us with a lot of presence data.

One step further into detail we have 'presence/absence data' where species lists of a certain location try to state all records of organisms, implicating those not mentioned being absent from the site. Depending on the origin of the research these presence/absence data can be restricted to a certain taxonomic level.

'Density data' go even further to provide an idea of the concentration, making it possible to have a better idea of the composition of present communities and species' assemblages.

Based on these previous three a fourth category could be created with 'virtual data', having its origin in *e.g.* habitat-suitability assessing methods (Hirzel *et al.*, 2001) or ecologically species response curves to environmental gradients (Austin, 2002).

In the four subsets of data mentioned, the two biggest sinks of data are the first and last category. While the last one can generate a lot of data, it also could play a role in directing our attention where to start for in the first category, at the same time testing the acceptability of the method used to generate the virtual data.

As there is a high dispersal rate of organisms in marine environments we should integrate ecological and environmental interactions. Most models consider the distribution of a species as the sum of a set of environmental constraints while two types of ecological interactions do play an important role to improve ecological models (Guisan *et al.*, 2006). First there are biotic interactions which can be interspecific (*e.g.* competition, predation, symbiosis) as well as intraspecific (*e.g.* competition, facilitation), second are predictor interactions where the effect varies depending on the other acting parameters, both biotic and abiotic. Experimental systems which are closed demographically thus do not reflect the real-life processes changing biodiversity, where - as well dispersal as immigration - play a crucial role in a marine ecosystem. This inherent species richness and variability can be translated to functional groups of allocated species based on several relevant characteristics; habitats with a different species composition can in this way contain the same functional groups. The resulting habitats will interact between two extremes, where minimal to no interaction results in an overall effect as the simple sum of all processes in each habitat, or where a habitat depends critically on biodiversity effects in others. A continuum is expected from very open to almost enclosed systems determined by hydrographical factors, life history and dispersal capabilities (Giller *et al.*, 2004).

Integration of ecological data

Probably every data centre complains in one way or another on the different formats scientists report their datasets. Still between the different data centres there is not (yet) a common interface to store or access all data. So why is it so difficult to have one single marine database able to store the results of our research and share them with the community? The answer is twofold: we drag our past along and we focus on what is interesting us. Our historical luggage decides often which system we are using and what kind of interface we use to manage this. Institutes happen to be run by people and even if there was no financial limit to hardware and software investments, research institutes use the in-house - human - experiences, so we keep at our job and we continue with choices made in the past. The creation of the conceptual model of a database based on a functional analysis is in this way the most important step in the development of a data management system. The structural differences of these systems, originates often in the scientific background of institutes maintaining them. The conceptual design to store data of a taxonomer or a geniticist won't have much in common. The challenge today is to interlink those different data systems in such a way the result has an added value. An ultimate benchmark in this interdisciplinary project is the gap between biologists and computer scientists who rarely understand the difficult conceptual problems of the other discipline (Guralnick & Neufeld, 2005).

As always there is no miraculous solution coping with all different possibilities. Datasets to be integrated can vary from a final spreadsheet to a relational database management system.

Importation of data is an option for those datasets typically 'closed' at the end of a project, often residing on a dark and forgotten spot on the hard disk of a scientist used to work for the project but ever since, has forgotten all about it.

Linking by URL's is a popular approach when the information is accessible through the web and maintained distantly. Keeping the link alive is the only data to maintain locally and if this is with the consent of the managers of the external site, all cobblers can stick with their last.

Warehousing data systems is a huge task, able to perform as well at database as at interface level or the combination of the two. Design of both (DB and UI) usually doesn't stand still and somehow, this structure is the origin of different locally maintained systems, so luckily this kind didn't catch on with the community.

The youngest option in this series and still a buzz word is connecting datasets by *web services* (Page, 2005). Basically it consists of two parts: a ‘metadata’ service that can be queried to retrieve the possible datasets and their formats (*e.g.* xsd schema) and a ‘data’ service providing the values (*e.g.* xml data) visualised through the web interface (*e.g.* xsl stylesheet).

With a similar structure, or a mix of these components, in place we are only half way to share data and make data truly accessible. From our experience in SeaDataNet (http://www.seadatanet.org/) we learned that ontology and mapping vocabularies is still an underestimated job. Data of true quality is valorised at this point. When for instance data are reported by parameters without a detailed description of the measurement procedure often these data can not be compared to those at other data centres (*e.g.* values of total nitrogen). Another well known ontology ‘avant la lettre’ is taxonomy. Naming a species obeys since Linnaeus to different rules, evolving ever since to rule books published by different international institutes, governing how organism should be named. The one goal is to get each organism a unique binomial name as its identification set the fundament to any biological study based upon it. ‘Errare humanum est’ and as a result we now have to deal with synonyms, homonyms, basionyms and differences in spelling or hierarchies (Chavan *et al.*, 2005).

Dealing with taxonomical data

From the start of the far shore windmill monitoring campaigns we understood our current database system wasn’t ready to deal with biological data (De Cauwer *et al.*, 2006). Looking around we were aware of different initiatives each one with its own focus (Sereno *et al.*, 2005) and often with a lot of manpower behind. Furthermore we didn’t want to create a new taxonomical standard to report biological data but otherwise a high level of hierarchy and different classifications should be stored.

In our quest for a taxonomic database we put forward some criteria to have a certain level of interoperability and continuity. Applicants should be an accepted standard to the scientific community, a dump of the database contents should be available at regular intervals and they have to be able to cope with variation in supraspecific taxa. It was this last one which ended up as the limiting factor and only two systems remained. Since our taxonomic tree has been online mid March page hits from external domains clearly shows the interest in hierarchical data storage and confirms this supposition.

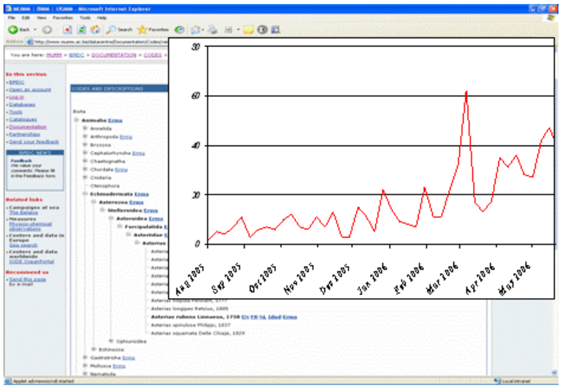


Fig. 2: Weekly hits on taxonomic tree past year

ITIS - Integrated Taxonomic Information System - (http://www.itis.usda.gov/) is one of these but as mentioned on there homepage “*Here you will find authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world.*” it contains mainly North American taxa. For this reason we have chosen its European counterpart ERMS - European Register for Marine Species - (http://www.marbef.org/data/erms.php) as our privileged reference. The connection between the datasets is based on the *linking* principle mentioned above with regular mass updates by *importing* updated taxon names and classifications.

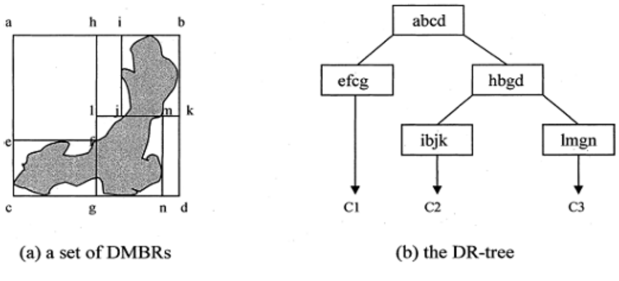
Using a structure like this has a lot of advantages while there are still a few drawbacks. As we do not have the specialists and the time nor a network of reviewers, like ITIS and ERMS, to track all taxonomic changes we have to outsource taxonomy. Still we need a hierarchical structure to store data as these are reported on different taxonomic levels. Moreover to be able to continue with the current trends in ecological modelling, the opportunity to create functional groups, species' assemblages and to define communities and habitats is a crucial prerequisite of a usable dataset. Moreover the two taxonomic interfaces have a broad taxonomic range while they are still referring to other more specialised web-interfaces and official publications thus providing in a way a portal to the taxonomic community. We see at the moment two main disadvantages to this system. First there is the big amount of taxa to be stored. As for each reported taxon the whole tree bottom-up is generated based upon data received by the originator or data retrieved from the web. This workload diminishes in time as bit by bit the whole taxonomic tree gets imported and as the dataset are available automatic procedures are being developed. Second disadvantage is the complexity of hierarchical querying. But also here we keep learning and future software updates promise extra functionalities to ease and speed up this specific kind of querying.

Spatial querying

Another important selection for biodiversity research is spatial querying. Georeferenced selecting systems now often used are bound to longitude latitude rectangles. Mapping geological parameters, species distributions and even national boundaries rarely follow these lines. With the increasing popularity of geographic information systems the last year big advances where made in spatial querying. As much of these new technologies do not concern marine science most of it falls out of our scope but still we can learn a lot from this evolution in spatial querying.

To define spatial objects there are for the moment two popular methods. First one is the object decomposition where the spatial object is recursively divided into smaller objects until the desired level of detail is attained. The second and most common used at the moment are based on the Minimum Bounding Rectangle (MBR) which is the smallest rectangle enclosing the spatial object. Spatial querying based on the latter is executed in two steps. The filter step limits possible candidates (called tuples) by using the predefined MBR's. As certain tuples can fall into an MBR but still be outside the spatial object (false positive hits) these filtered tuples are passed on to a refinement step where the exact representation is checked based on algorithms.

The R-tree-family spatial index structure is based on this principle and surfaces frequently in spatial database literature. For the moment we focus on the performance of the filter step, as limiting the number of false positive hits, limits CPU time to run sometimes complex algorithms. Two practices are rather easy to incorporate into a database and we shortly will present these. Decomposed MBR is the MBR of a spatial object which is recursively decomposed. The resulting DR-tree should be stored at a leaf-node of the R-tree (Lee & Chung, 2000).



(a) a set of DMBRs

(b) the DR-tree

Fig. 3: Representation of a DR-tree (Lee & Chung, 2000)

The second way is to store indirect predicates of child nodes at the domain level to make it possible to recognise false hits at an earlier level en in this way limiting the number of candidate tuples at an early stage of the filter step (Park, 2004).

Depending of the level of detail we set to the spatial querying and the error estimates about the observations it could still be a possibility to return the candidate tuples to the user without executing the refinement step. In this scenario we would gain a lot of performance and let the final decision whether to use data or not to the end user. Moreover, spatial objects defined by our web interface will never be very complex so even if the refinement step is executed the overall performance should be acceptable.

Conclusion

As a marine data centre we are used to deal with big datasets as data acquisition in marine science has been automated the last decade. Contrarily biological data is for its biggest part manual and the binomial classification used to identify organisms wasn’t clearly created during the information age. As those identifications are a corner stone for ecological modelling spatial joining will soon be a prerequisite for usable data. As these defined species communities will depend on spatial distributions linked to physical, chemical and geological parameters it is clear we have to cope with complex spatial object to query biological data in our database.

The ease to use our existing database structure and the possibility to incorporate new functionalities to handle this ecosystem approach in the context of the monitoring of human impacts at the North-Sea clearly demonstrates the advantages of such a logical design. The ongoing developments and the need to access qualified data only grow stronger. As data centre we therefore have to keep in mind our job is to archive marine data but we should also put ourselves at the service of the scientific community and provide data and applications in a comprehensive format to be used at an international level.

Acknowledgements

The achievements of this integrated data system and its applications are partly funded by the Belgian Science Policy Office.

References

Austin, M.P. (2002), Spatial prediction of species distribution: an interface between ecological theory and statistical modelling, Ecological Modelling 157, 101-118

Chavan, V., Nilesh, R., Watve, A. & Ruggiero, M. (2005), Resolving taxonomic discrepancies: role of electronic catalogues of known organisms, Biodiversity Informatics 2, 70-78

De Cauwer, K., Devolder, M., Jans, S., Meerhaeghe, A. & Scory, S. (2006), IDOD: an integrated marine and environmental database, Proceedings of the 2006 ICES Annual Science Conference, Maastricht September 19-23 2006

Giller, P.S., Hillebrand, H., Berninger, U.G., Gessner, M.O., Hawkins, S., Inchausti, P., Inglis, C., Leslie, H., Malmqvist, B., Monaghan, M.T., Morin, P.J., O'Mullan, G. (2004), Biodiversity effects on ecosystem functioning: emerging issues and their experimental test in aquatic environments, Oikos 104 (3), 423-436

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.MC., Aspinall, R. & Hastie, T. (2006), Making better biogeographical predictions of species' distributions, Journal of Applied Ecology 43, 386-392

Guralnick, R. & Neufeld, D. (2005), Challenges building online GIS services to support global biodiversity mapping and analysis: lessons from the mountain and plains database and informatics project, Biodiversity Informatics 2, 56-69

Hirzel, A.H., Helfer, V. & Metral, F. (2001), Assessing habitat-suitability models with a virtual species, Ecological Modelling 145, 111-121

Lee, Y.-J. & Chung, C.-W. (2000), Analysis of two-step index structure for complex spatial objects, Information Sciences 125, 133-152

Maes, F., Schrijvers, J. & Vanhulle, A. (red.) (2005), A Flood of Space, Belgian Science Policy, Brussels, 204 pp.

Page, R.D.M. (2005), A Taxonomic Search Engine: Federating taxonomic database using web services, BMC Bioinformatics 6:48

Park, H.-H., Min, J.-K., Chung, C.-W. & Chang, T.-G. (2004), Multi-way R-tree joins using indirect predicates, Information and Software Technology 46, 739-751

Raffaeli, D. (2006), Biodiversity and ecosystem functionality: issues of scale and trophic complexity, Marine Ecology Progress Series 311, 285-294

Sereno P.C., McAllister, S. & Brusatte, S.L. (2005), TaxonSearch: a relational database for suprageneric taxa and phylogenetic definitions, PhyloInformatics 8, 1-21