CM2006/M:23

IDOD: an integrated marine environmental database

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

Abstract

In 1997, the Belgian Marine Data Centre started developing an integrated marine environmental database. The initial core model was designed taking into account the characteristics of concentration data in seawater. This model was soon expanded to cope with concentrations in air, sediment and biota. Nowadays, also spectral data (optical properties, sediment granulometry), observations on living organisms and biodiversity information are stored in the database. Very recently, the structure was complemented with a taxonomic module that allows the storage of vernacular names, taxonomic changes, determinations to any level of detail as well as revised determinations of a given specimen. Next to the storage of measurement results, special attention is paid to the storage of a comprehensive set of meta-information like the corresponding research projects, the methodology and the related quality control information. The main thought about storing and linking results of various disciplines together is to improve the knowledge of our marine ecosystem in its broadest point of view and to make the life of the users of the system easier. The online interface provides many search options and several export formats, in order to allow everyone to create a subset of the data suited to his/her very own needs. In this paper, we present the structure of the IDOD database, the technical choices we made to reach our goals and our plans for the future.

1. Introduction

The Management Unit of the North Sea Mathematical Models (MUMM) is the Belgian federal scientific service responsible for the protection of the marine environment. It has since its creation in 1976 always been involved in data collection and data management. At the end of the 90’s, the need for a new relational database management system became evident. Thanks to the funding of the Belgian Federal Science Policy, the IDOD project was then launched to cover all the data management needs of both the structural monitoring programmes performed by and on behalf of the State and the various research projects performed in the universities and similar scientific institutions.

An increasing variety of datasets needed to be stored, archived and disseminated to the scientists, the policy makers and the general public. The data and/or metadata of the monitoring and research efforts needed to be reported to international organisations and networks according to predefined formats (*e.g.* ICES, European Directory of Marine Environmental Datasets (EDMED) …). Since 1998, an integrated marine environmental database with supportive tools was developed. In this database different data types, like spectral data, biodiversity information, observations and concentrations in seawater, air, sediment and biota, are well documented and integrated in a sole structure. By integrating the different dissemination and reporting functions as well, the information system facilitates greatly the data managers work.

2. Methodology

The most important and first phase of database design is the creation of the conceptual model describing the information to be stored as abstract objects and their associations. With the full integration of all data types into one model in mind, we addressed first the case of the concentrations of substances and similar parameters in sea water. Once that model had been validated, it was expanded to concentrations in air, sediment and biota. Afterwards, spectral data, observations and biodiversity information were added.

Several bilateral contacts with different data providers were necessary to understand the characteristics and to foresee the import of their data sets. The pan-European data catalogues EDMED, European Directory of Marine Environmental Research Projects (EDMERP), CSR and the ICES environmental reporting format were taken into account to ensure the exchangeability of our holdings in the required format.

All concepts to be stored were carefully described:

- each entity with all its attributes (field type, format, optionality, ...) and a primary key,
- each relation with its cardinalities (minimum and maximum occurrences that can exist for a relation from an entity to another) and its transferability ("is the relation updatable?"),
- an 'Entity – Relation' diagram,
- the business rules to maintain the quality and consistency of the data.

For most parts, the three first normal forms of the database conception rules were checked. Normalization is a process that eliminates redundancy, organizes data efficiently, reduces the potential for anomalies during data operations and improves data consistency. Edgar F. Codd (1970) originally established three normal forms, later refinements were made as additional normal forms. Normalization is described in more detail by Morris (2005) and for example on Wikipedia (2006).

The main objective, the storage of samples and results, was kept in mind to avoid an overnormalisation and complex design requiring a lot of programming effort. As such, the structure for e.g. publications and authors was not fully normalized. Whenever this part would gain in importance, the structure can still be adapted.

During the logical design phase, the functional model is converted into a technical design. The relations between entities are translated in foreign keys, or transition tables in case of many-to-many tables, the indexes to be created were reviewed and the triggers and procedures were written. The resulting table definitions can directly be implemented as 'physical' tables by executing the data definition queries. The CASE tool Designer/2000 was used to construct the conceptual model, to convert this model into table definitions and to build the data definition SQL code. All documentation of the system, including the change history, is stored in the CASE tool.

3. The conceptual model

3.1. General structure

A general view of the resulting structure is given in Figure 1. The central part of the scheme represents the physical samples, grouped by sampling event (same place and time). They are directly linked to the values for water and biodiversity measurements and divided into subsamples in the case of sediment and biota samples. The spectral values are related to samples via the entity SPECTRA.

Samples are collected in the frame of a “project”, during an oceanographic “campaign” on a “platform”. The platform can be a research vessel, "on foot", an automatic measurement station, *etc*. Specification of the involved services and persons are important for both the campaign and the project. The field ‘reference sampling depth’ (*e.g*. ‘bottom’, ‘subsurface’…) is added to allow the export of a tabular format in which values of different samples can be grouped on same place, time and depth.

SAMPLES are thus characterised by a position, sampling depth, time and sampling method. The position can either be an exact position or the reference position of the sampling station to which it is linked. In case of tracks, the start and end positions and times are entered. A flag indicates when the time reference system is not known (otherwise converted to UTC). Similarly, another flag states when the position reference system is unknown.

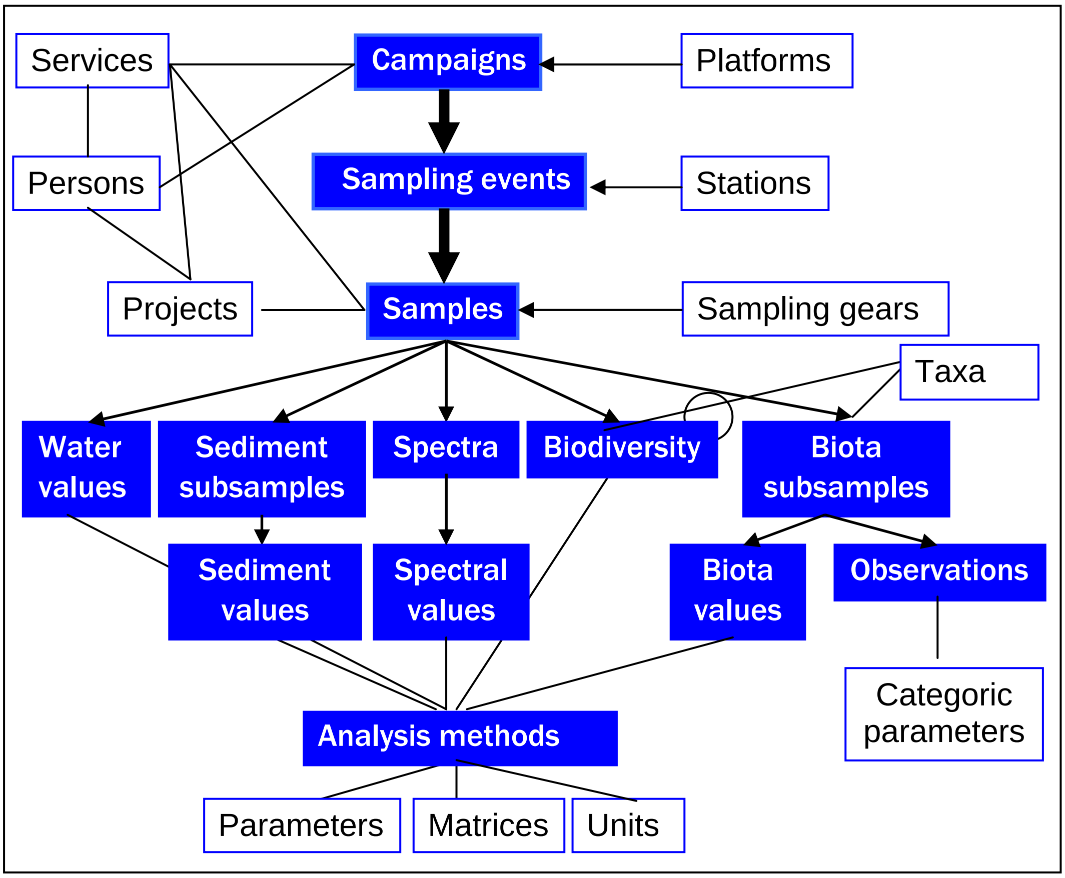


Figure 1. General structure of IDOD database

As parameters measured on water and air samples have the same characteristics, those results are stored in the same entity WATER VALUES. Sampling depths are negative for air and positive for water and the matrices differ: ‘rainwater’, ‘aerosols’ or ‘gas’ for air; and ‘dissolved’, ‘particulate’ or ‘total’ for water.

The optical data like total backscattering coefficient are characterized by a range of values at given wavelengths. They are grouped by SPECTRA, the entity defining the analysis method, project and processing flag.

For sediment data, the entity SEDIMENT SUBSAMPLES was created for the storage of depth profile data. Extra attributes are necessary to define the fraction sizes of the sediment analyzed.

A sample of living organisms is often subdivided in individuals or pooled subsamples of a given species or species group (defined by TAXA, see section 3.2). These BIOTA SUBSAMPLES can be characterized by physical characteristics like weight and length. Numerical measurement results like contaminant concentrations on a specific tissue of the subsample are stored as BIOTA VALUES while OBSERVATIONS contain the so-called ‘categorical values’. OBSERVATIONS contain any data like oil coverage in %, presence of parasites or cachexy index for birds of a given subsample of specimen(s).

BIODIVERSITY measurements, like bird counts, phytoplankton biomass and mammal abundance, are directly linked to the sample and to a given species or species group (TAXA). The same specimen can be determined several times by several people over the years. Especially in the case of historic collections, redeterminations and taxonomic revisions are very important. Actually, the BMDC is involved in the digitization process of the Gilson collection. These samples, collected by G. Gilson in the Southern North Sea between 1899 and 1910 are stored at the Royal Belgian Institute of Natural Sciences. The number of determinations by different persons during the last century varies from specimen to specimen. By linking the biodiversity results to themselves, any number of determinations of a specimen can be recorded with specification of date and observer.

All resulting values are linked to an ANALYSIS METHOD, that describes and links to all information on the analysis, sample handling, parameter, matrix, unit and analysis laboratory. To incorporate information on the quality of the analysis method, two different aspects are introduced: intercomparison exercises and control chart information.

3.2. Taxonomic structure

At the early stage of the development process of the database, a simple denormalized structure for information on species was generated to link with the results of densities and biomass of given species, observations on species and contaminant concentrations in specimens. Although that initial structure was useful for species with a well-established taxonomy and determinations to the same level, it proved to be insufficient to cope with all the datasets we had to handle.

In practice, datasets pertaining to living resources range from contaminant concentrations in well-known commercial fish species to densities of a dynamic taxonomic group like nematodes. In the latter case, the problem is even more complex because determination is not always done at the species-level. New species are encountered and reported before they have an official scientific name. Also some benthic faunal groups were determined to ecological community level, with for each community a list of species, subspecies and genera of which it consists.

A flexible system was needed in order to deal with the hierarchal nature of the data, like in taxonomic databases (European Register of Marine Species (Costello, 2004), (Costello, 2000), Integrated Taxonomic Information System (ITIS, 2005)). Also identifiers of taxa in international taxonomic databases, synonyms and informal names needed to be stored. Informal names are interpreted as those names which are not part of the official taxonomic vocabulary like temporary names given by biologists, names in different languages and commonly used names like zooplankton.

In the entity TAXON (see figure 2), the scientific taxon name at a given level (link with TAXON_LEVEL) is entered and the mandatory link with the parent as a link to itself. This relation is one-to-many: each taxon has one and only one parent, each taxon has zero to many child taxa. Besides the scientific names, also non-official names or 'clades' are entered in this entity and the relation with the scientific names is established via the intermediate table TAXON_CLADE to reflect the many-to-many relation. The term 'clade' is used in our database to define all informal names. Clades and taxa are stored in the same table. Only the informal names are linked to CLADE_SCHOOL. CLADE_SCHOOL defines the type of informal name and has entries like 'English', 'ERMS ID', 'Euring' or 'name given by person x of service x'. This way a very flexible system is obtained that is able to deal with the uncertainty in determinations and taxonomy. The results are stored in the way they are reported by the data originators.

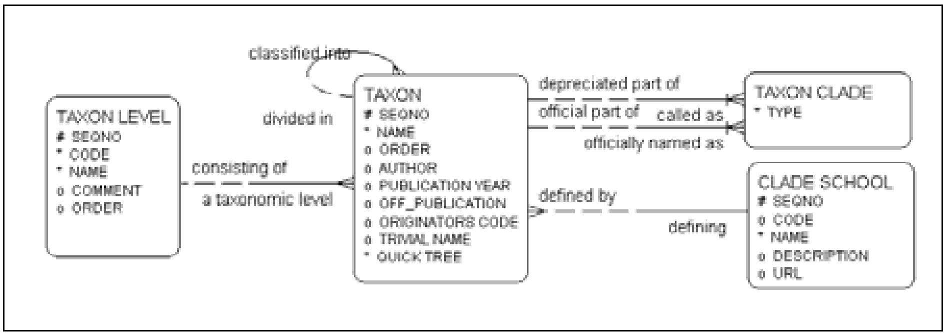


Figure 2. E-R diagram for taxonomy.

For example, in some cases it is hard to distinguish two or more species. Data providers report the results with the name of both species, *e.g.* zooplankton species *Chaetoceros compressus-costatus.* In the IDOD database the results will be linked to the non-official name *Chaetoceros compressus-costatus* that is entered in the table TAXA. The individual species are entered as well and the link between the informal name and the two official species is made via TAXON_CLADE. The clade school, linked to the informal name *Chaetoceros compressus-costatus,* defines who has given the name.

Another example is the determinations of Copepoda to a community level like Copepoda epibenthic species. The densities are directly linked to the group of species while this group is linked to the official species of which it may consist.

Also the taxonomic discrepancies as described by Chavan et al. (2005) like differences in taxonomic hierarchies, differences in spelling and homonyms can be imported as 'clades' in this structure.

Although the structure is capable of, it is not the aim of the Belgian Marine Data Centre to store the complete taxonomic tree. Instead, only taxa for which measurements or observations exist are entered in the database. Reference is made to ERMS (European Register of Marine Species) and ITIS (Integrated Taxonomic Information System) for the complete and most recent taxonomic information.

3.3. Structure to handle user access rights

Users are able to select and download data online. With regards to the data policy, users do not have the same rights on all data sets. The practical data policy rules can also depend on the project for which the data are gathered.

For instance, results of official monitoring campaigns are available for the public immediately after their incorporation in the data base. For the research projects in the frame of national research programmes, an embargo of two years after the data transmission deadline is foreseen. After the embargo all data can be used by anyone respecting the data access rules. Before the embargo, the data are private and only available to persons nominated by the data originator. Besides this, a few completely private datasets also exist in the database.

To respect all these rules, two fields in the table PROJECT are added and a table SPECIAL ACCESS was created as shown in Figure 3. A project can be flagged as either public or private. For private projects an embargo period can be entered after which the data become public. The persons granted access to data of a given project, even when a project is flagged as private, are linked to that project via the table SPECIAL ACCESS. Business rules ensure that data are only returned when allowed: when a query for data is executed by a given user, the project of every value is checked. When the project is flagged as private and the embargo period is not passed, a procedure checks if the person is granted access for this project. When the user has access, the values are returned, if not, only the existence and number of private values are mentioned.

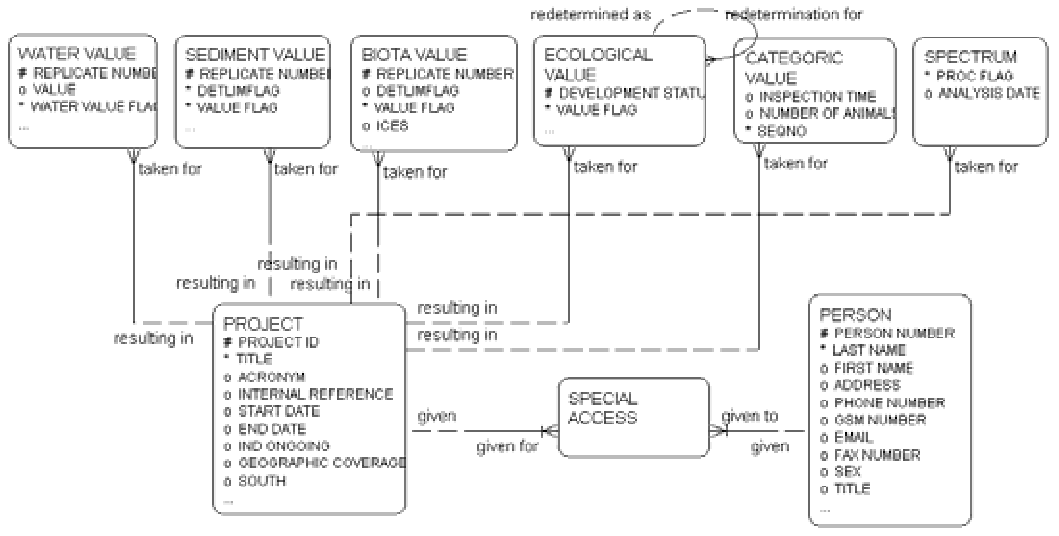


Figure 3. E-R diagram for user access.

All requests performed are stored in the table REQUESTS. The user can recall an existing request to modify it or to retrieve newly imported data meeting the same selection criteria, if any. The table is also used to log the data selection activity.

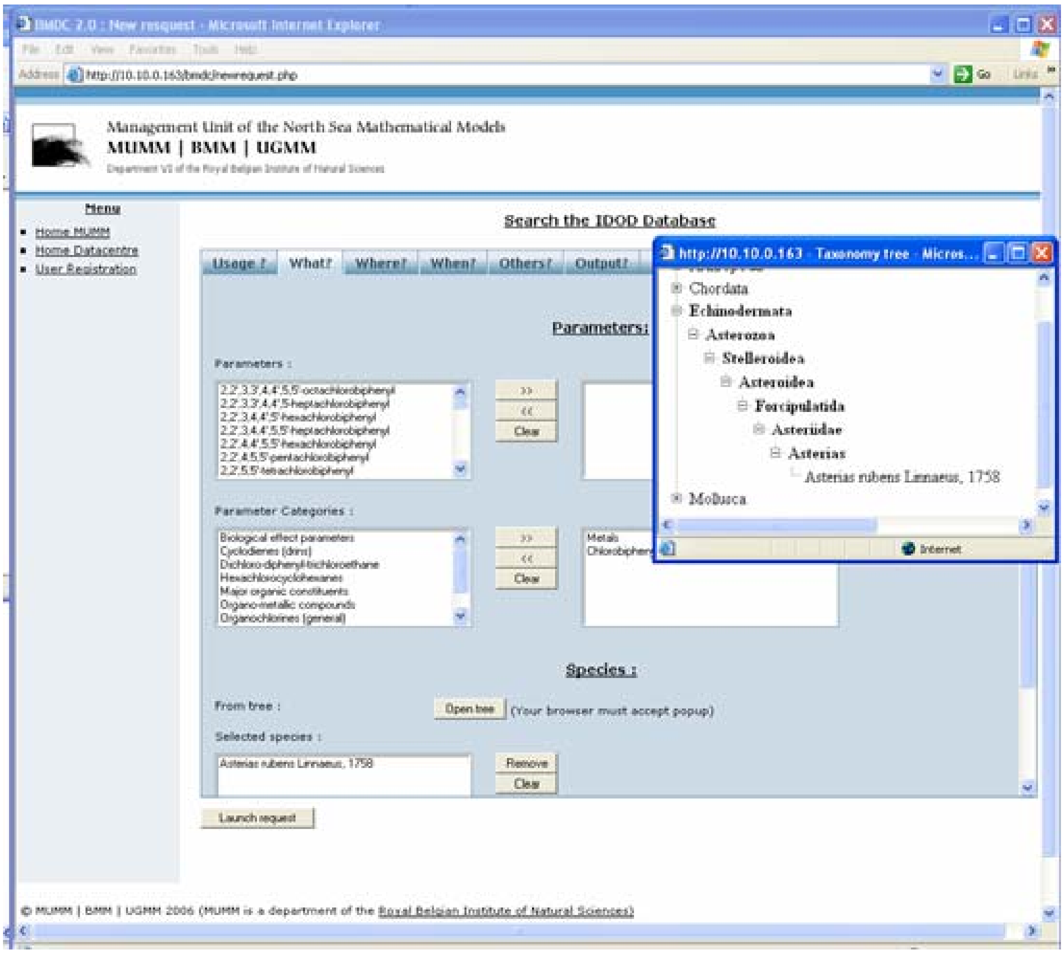
4. Database

The resulting database (Oracle8i) consists of more or less 115 tables of which 39 reference tables and 19 tables used for importing and reporting. Business rules are implemented as PL/SQL procedures and triggers. Triggers are used on every table to record auditing information like by whom and when the record is created and updated. For the table SAMPLES and REQUESTS a complete journaling table is kept. For every change (update, delete, insert), a record is stored with the user name and date of change. Imports of data are usually done with SQL commands and triggers on import tables verifying data integrity and consistency.

Migration to Oracle 10g is foreseen in the near future. Java was used to create the first version of the user interface. The new interface (see next section) is written in the scripting language php.

5. User interface

An online user interface has been created to enable users to select and download data after registration. The requested data type should be selected first. In a next step, several criteria, grouped in tab pages, can then be specified. Species or higher levels can be selected in a taxonomical tree or can be entered as free text. Geographical selection on a map will be available in the near future (Meerhaeghe *et al.*, 2006). The users are able to edit and launch their previously made requests. The resulting dataset is returned in record format with one line per value. Optionally, also a tabular format can be selected as output format, in which the data are presented in columns per parameter or parameter-method suitable for data analysis. All documentation like the lists of codes used, explanation of the output format and an inventory of the data available in the IDOD database is presented on the web site.


BMDC 2.0 : New request - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://10.10.0.163/bmdc/newrequest.php
Go
Links
Management Unit of the North Sea Mathematical Models
MUMM | BMM | UGMM
Department VI of the Royal Belgian Institute of Natural Sciences
Menu
Home MUMM
Home Datacentre
User Registration
Search the IDOD Database
Usage ?
What?
Where?
When?
Others?
Output?
Parameters:
>>
<<
Clear
>>
<<
Clear
Species :
Open tree
Remove
Clear
Launch request
http://10.10.0.163 - Taxonomy tree - Micros...
Internet

Parameters :

2,2',3,3',4,4',5,5'-octachlorobiphenyl
2,2',3,3',4,4',5-heptachlorobiphenyl
2,2',3,4,4',5'-hexachlorobiphenyl
2,2',3,4,4',5,5'-heptachlorobiphenyl
2,2',4,4',5,5'-hexachlorobiphenyl
2,2',4,5,5'-pentachlorobiphenyl
2,2',5,5'-tetrachlorobiphenyl

Parameter Categories :

Biological effect parameters
Cyclodienes (drins)
Dichloro-diphenyl-trichloroethane
Hexachlorocyclohexanes
Major organic constituents
Organo-metallic compounds
Organochlorines (general)

Metals
Chlorobiphen

Chordata
Echinodermata
Asterozoa
Stelleroidea
Asteroidea
Forcipulatida
Asteriidae
Asterias
Asterias rubens Linnaeus, 1758
Mollusca

From tree : (Your browser must accept popup)

Selected species :

Asterias rubens Linnaeus, 1758

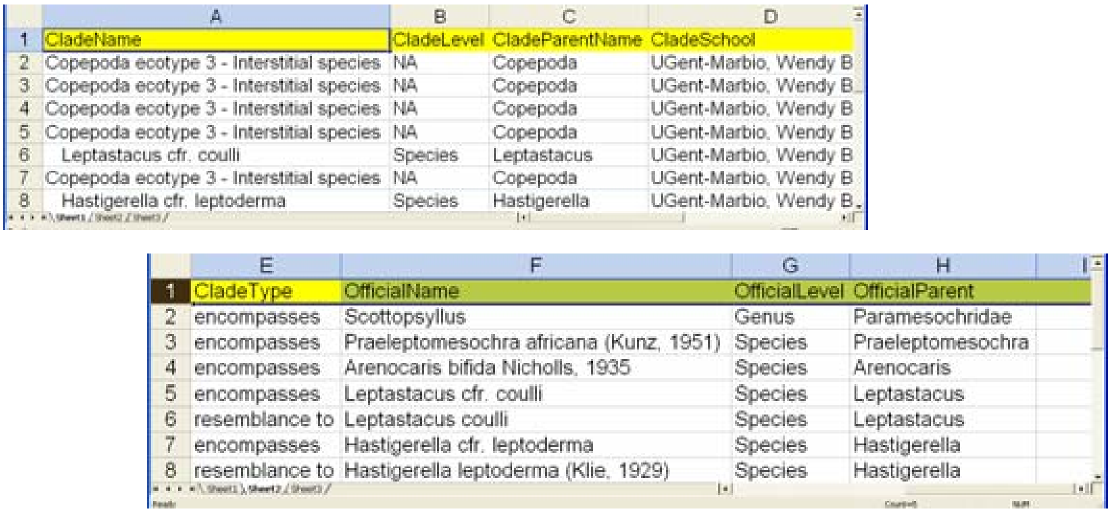
© MUMM | BMM | UGMM 2006 (MUMM is a department of the Royal Belgian Institute of Natural Sciences)

Figure 4. User interface.

From the developer’s point of view, a disadvantage of the highly normalized taxonomic structure is the quite complex SQL codes necessary to visualize the taxonomic tree or to retrieve data of a given taxon. An advantage of OracleSQL is the support for querying hierarchical data with the START WITH and CONNECT BY clauses allowing a hierarchical presentation of the data (Gennick, 2003). Other database systems would need a code that recursively issues a series of SQL statements to walk up or down the tree (Morris, P.J., 2005).

The user, however, can be sure that all results that are of possible interest to him/her are returned. When a taxon of any level is selected in the query interface, SQL code ensures that not only results for that taxon are returned but also for all ‘child’ taxa, synonyms of the selected taxon and informal names linked to the selected taxon. For example, when *Chaetoceros compressus* Lauder 1864 is selected in the tree on the query interface, the results for the non-official name *Chaetoceros compressus-costatus* are also returned. Together with the resulting text file, a file is returned with information on the non-official names used.

In Figure 5 an example of such a file is given for ‘Copepoda communities’. The ‘Copepoda ecotype 3 – Interstitial species’ community, accompanied with information about its level, parent and clade school, encompasses a list of different official species and genera. If however an informal species name is encountered in this list, *e.g. Leptastacus cfr. coulli* in line 5, this name returns indented, on the next line in the column ‘clade’ resembling the official species *Leptastacus coulli*.



	A	B	C	D
1	CladeName	CladeLevel	CladeParentName	CladeSchool
2	Copepoda ecotype 3 - Interstitial species	NA	Copepoda	UGent-Marbio, Wendy B
3	Copepoda ecotype 3 - Interstitial species	NA	Copepoda	UGent-Marbio, Wendy B
4	Copepoda ecotype 3 - Interstitial species	NA	Copepoda	UGent-Marbio, Wendy B
5	Copepoda ecotype 3 - Interstitial species	NA	Copepoda	UGent-Marbio, Wendy B
6	Leptastacus cfr. coulli	Species	Leptastacus	UGent-Marbio, Wendy B
7	Copepoda ecotype 3 - Interstitial species	NA	Copepoda	UGent-Marbio, Wendy B
8	Hastigerella cfr. leptoderma	Species	Hastigerella	UGent-Marbio, Wendy B

	E	F	G	H
1	CladeType	OfficialName	OfficialLevel	OfficialParent
2	encompasses	Scottopsyllus	Genus	Paramesochridae
3	encompasses	Praeleptomesochra africana (Kunz, 1951)	Species	Praeleptomesochra
4	encompasses	Arenocaris bifida Nicholls, 1935	Species	Arenocaris
5	encompasses	Leptastacus cfr. coulli	Species	Leptastacus
6	resemblance to	Leptastacus coulli	Species	Leptastacus
7	encompasses	Hastigerella cfr. leptoderma	Species	Hastigerella
8	resemblance to	Hastigerella leptoderma (Klie, 1929)	Species	Hastigerella

Figure 5. Example of file with information on the non-official names used.

With regards to the redeterminations of biodiversity data, the number of previous determinations is given as an additional column in the results file. Only the most recent determination is returned in the file. All previous determinations, if any, can be requested by mail.

6. Discussion

The IDOD database is constructed for the long-term archival of a large range of different data types. All marine analytical results and observations are stored with the corresponding meta-information necessary for a good understanding and interpretation of the results. In the IDOD database, the results related to living resources are linked to a hierarchical taxonomic structure that also enables the storage of vernacular names and revisions of taxonomy. An online user interface was created for the access to the different data types. In the query interface, different criteria can be selected as well as a tabular output format besides the usual record format.

We can not forget that the usefulness of such an information system is determined by its contents (Diepenbroek *et al*., 2002). Table 1 shows the number of data actually available in the database. Taking into account that the import of data started only mid-2001 by a small team (2.5 full time equivalents) also responsible for developments, and the number of datasets in the ‘import’-queue, these figures are really encouraging.

Table 1. Contents of the IDOD database.

Table	Nr of records
Projects	252
Campaigns	695

Samples	54942
Parameters	393
Methods	668
Data items	324389
Taxonomic items	4211

Even though design and developments costs are high, the data managers, data users and moreover the quality of the data benefit from the system. It is the central tool used daily for a range of data management functions. Since its opening in July 2002, around 340 persons registered to use the database and 800 requests were executed with, for every request, almost two runs: with or without modified selection criteria.

New developments are continuously ongoing to deal with the external changes, to improve the performance and to broaden the range of data types. It has always been possible to extend the model without major problems. As a next step, an indexing and inventory system for large data files like multibeam records, video or seismic data is foreseen.

Acknowledgements

The ongoing development of the IDOD information system is partly funded by the Belgian federal Science Policy Office.

References

Chavan, V., Rane, N., Watve, A. (2005). Resolving taxonomic discrepancies: role of electronic catalogues of known organisms. Biodiversity Informatics 2 (2005) 70-78.

Codd, E.F. (1970), "A Relational Model of Data for Large Shared Data Banks", Eprint, *Communications of the ACM*, 13(6) 377–387.

Costello, M.J.; Bouchet, P.; Boxshall, G.; Emblow, C.; Vanden Berghe, E. (2004). European Register of Marine Species. http://www.marbef.org/data/erms.php (28-Jul-2006).

Costello,M.J. (2000). Developing species information systems: the European Register of Marine Species (ERMS). Oceanography Vol. 13 No. 3/2000.

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R. and Wefer, G. (2002). PANGAEA – an information system for environmental sciences. Comuters & Geosciences 28 (2002) 1201-1210.

Gennick, J. (2003). Querying Hierarchies: Top-of-the-Line Support. http://www.oracle.com/technology/oramag/webcolumns/2003/techarticles/gennick_connectby.html (28-Jul-2006).

ITIS (2004). About ITIS. Standards and Database documentation. http://www.itis.usda.gov/standard.html (28-Jul-2006).

ITIS (2005). Integrated Taxonomic Information System. http://www.itis.usda.gov/ (28-Jul-2006).

Meerhaeghe, A., De Cauwer, K., Devolder, M., Jans, S. and Scory, S. (2006). Revealing species communities in a spatial and temporal overlap. ICES Annual Science Conference 2006, Maastricht.

Morris, P.J (2005). Relational Database Design and Implementation for Biodiversity Informatics. PhyloInformatics 7 (2005) 1-66.

Wikipedia (2006). http://en.wikipedia.org/wiki/Database_normalization (28-Jul-2006).