IDOD : Integrated and Dynamical Oceanographic Data management

Leader :

G. Pichot, K. De Cauwer, M. Devolder, S. Jans, M. Moens, L . Schwind, S. Scory

Partners :

J. Van Dyck, B. Plevoets, G. Dierckx

J.-P. Donnay, M. Binard, Y. Cornet, F. Muller

RBINS - Management Unit of the Mathematical Models of the North Sea
KULeuven - University Centre for Statistics
ULg - Laboratory SURFACES

The existence of a structured and validated knowledge base is an obvious need for any scientific work, especially when dealing with the marine environment. Any policy to be defined or decision to be taken in the perspective of a sustainable management of the North Sea would be meaningless without a background of validated and readily accessible measurements or experimental data. The set–up of an integrated oceanographic database was thus a key action in the frame of the "Sustainable management of the North Sea" programme.

The IDOD project was meant to establish, to manage and to promote a data base of marine environmental data, ensuring a smooth and scientifically sound data flow between the data producers (routine monitoring, field and laboratory experiments, mathematical models, ...) and the end users (scientists, sea professionals, policy makers, ...).

More specifically, IDOD would play the role of the "Programme Data Manager" for the measurements and data collected within the frame of the programme "Sustainable Management of the North Sea".

The project was split into five different –but highly inter–dependent– tasks:

As a basis, an inventory of the relevant data sets and databases was undertaken, in order to make them ready for incorporation in the database (a. o. with respect to current standards on data quality and on data documentation).

The procedures pertaining to the incoming flow of data were defined and implemented. This covers not only the practical aspects of the transfer of information but also the very important point of data quality control.

The design of the data base itself has been deeply analysed in function of the intrinsic characteristics of the data and in order to meet the present and future needs, ensuring the viability and the usefulness of the tool over the years.

In order to understand the processes driving the marine phenomena "hidden" in the data, a set of data analysis tools have been developed and are now tuned. Various approaches were used: statistical techniques, geostatistics and spatial analysis, space and time "corrections" of data sets by means of advection-diffusion models. Part of the information given by these tools is also used to improve the quality control on the incoming data.

Finally, as one of the most important objective of this project was to provide useful and scientifically sound information to a wide range of users, derived products (maps, tables, reports, ...) that meet the specific requirements and level of expertise of the various categories of users were designed and currently being made available to the users.

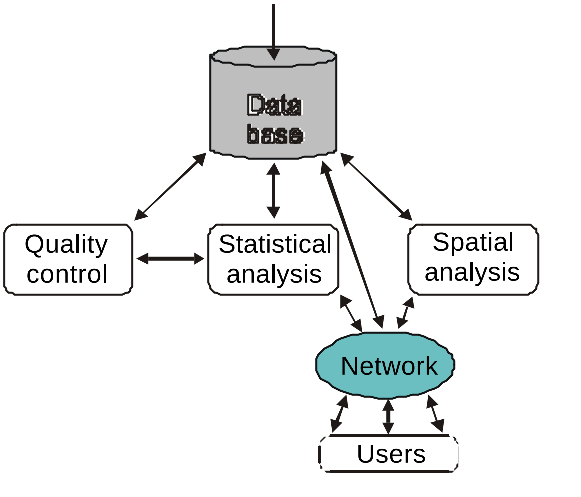
The global methodology applied to reach our objectives reduces to the following words: analysis, design, implementation and production.

The *analysis* phase went into the details of the data, their structure and intrinsic characteristics, together with a deep insight into the sampling methods, the laboratory practices and the needs and requirements of the potential users.

During this process a special attention was paid to the rights of the parties involved: the data centre, the data producers and the financing authority. After some negotiations, a formal convention specifying the 'rights and duties' of each of the parties could be agreed.

In parallel, in order to improve the consistency of the meta–information bound to the data an automatic recording system has been developed. It is currently being tested on board the *R/V Belgica*. It helps scientists to automatically keep track of important information such as date, time, position and environmental conditions for every sampling.

During the *design* phase the results of the analysis were translated into functional description: *how would data be entered*?, *how would their quality be checked*?, *how would they be retrieved and analysed*?, ..., leading finally to the *implementation* of the information system.



Sketch of the IDOD information system

The resulting system, now entering its ‘production’ phase, consists of a relational data base (running under *Oracle 8*i), quality and statistical analysis tools (based on *SPlus*) and visualisation and spatial analysis tools (developed with the help of *ArcView* and other *ESRI* software packages). As far as the technique makes it possible, processed data are made available to the users via the Web.

Content

The database mainly contains values of the concentrations of numerous substances in the air, the water, the sediment and the biota. These values result from measurements taken in situ and analyses carried out in laboratories. In addition to the concentrations, quantitative (biodiversity) and qualitative (pathology) information on the biota is also stored.

These values would be pointless if they were not accompanied by precise information about the circumstances in which they were measured. This is what is known as 'meta-information', a term that covers information such as the position in which samples were taken, the date, the time, the weather conditions, the sampling and analysis methods used, etc.

The database already contains several tens of thousands of items. All these data, documented and verified, constitute a coherent and unique source of information for scientists and other users.

Global achievements with respect to the objectives of the Programme

The partners of the project have the feeling that the information system they have built fully meets the priorities and objectives of the Programme ‘Sustainable Management of the North Sea’.

The categories of data that are considered cover a wide range of *natural processes* and of *human activities* pertaining to the North Sea. Up–to–date *quality control* standards are applied to the data being incorporated into the data base as well as to the processing procedures. The tools and products made available allow to *better understand the structure and the working of the marine ecosystem* and the influence of the human activities thereon, by providing a basis for *scientific assessments in the perspective of the definition of a sustainable management policy* of the North Sea. This is of primary interest for the Belgian obligations in the frame of the *International conferences on the protection of the North Sea* and of the Oslo and Paris conventions.

In the broader scope of the “Scientific support plan for a sustainable development policy”, our achievements also meet the stated objectives by providing *scientific tools* (the data base itself, the post-processing tools and the derived products) *that facilitate the process of decision making*. The development of the system and its feeding by the data collected in the course of the Programme have improved *the co–operation between marine scientists* and will offer them new ways to *improve their visibility and co–operative contacts at the international level*.

Future

The IDOD project has given the scientific community, the Belgian authorities and other potential users the opportunity to dispose of an up–to–date information and management system about the quality of the marine environment. The sea evolves continuously as do the needs and the demands of the society. The data base, together with its query and analysis tools, opens the door to new scientific investigations and to better policy choices. The onus lies on all the interested parties –the DB managers, the data producers, the authorities and the community of users– to keep this knowledge base useful, *i. e.* alive and up–to–date.

– – – – – – –

STATISTICAL ANALYSIS

The practical value of a database crucially depends on the quality of the data contained in the database and the ability to analyse these data in an intelligent manner. The University Centre of Statistics of the KULeuven therefore developed two supporting tools:

1. a *Statistical Quality Program* (SQC) that allows to set-up and apply in a systematic manner quality checks for the different data types. This tool is a stand-alone program that is to be used by the data manager to “qualify” new data that enter into the database;
2. a *Statistical Analysis Tool* (SAT) that allows to apply the most common statistical functions to a data set contained in the database. This second tool is available to any user of the database and runs on the world-wide web.

Because of the wide variety of data types in the database and because the contents of the database may dynamically change, the SQC program has been set-up as a general user-friendly “shell” that organises the general aspects of the quality checking in a consistent manner by supporting following functionalities:

- declaration of new data types. A data type consists of a list of measurement types that are concurrently measured;
- declaration of a test scheme for a given data type. A test scheme consists of a sequence of test stages, where in each stage a sequence of individual tests are applied to one or more of the measurement types held within the data type;
- support for following 3 generic types of individual tests: 1. a distribution test where the data value is checked against the marginal distribution which can possibly vary as a function of covariate variables (i.e. season, location,...), 2. a regression test where the data value is checked against the expected value obtained from concurrent measurements (in this case the subset of data that leads to the best regression is used); 3. a variogram test where the data value is checked against the nearest measurement of the same type within a given time-window;
- application of a test scheme. In such an application, measurement specific parameters are automatically retrieved from the buffer where statistical results (*i.e.* obtained through application of SAT) are maintained. The individual test results are combined through Bayesian analysis in a single numeric quality indicator varying from 0 (“statistically consistent”) to 1 (“statistically perfectly consistent”). A special feature of this method is that tests that cannot be applied (*i.e.* because data are missing) do not jeopardise the application of the test scheme.

The SQC program thus allows to statistically validate the data and report the results in a methodologically consistent manner, while remaining sufficiently flexible to adjust the validation to the different data types and update the test parameters as more information is obtained through statistical analysis of the data in the database.

The SAT program is a user-friendly tool that runs on the web and allows to perform standard statistical analyses: *i.e.* graphical and numerical statistical summaries, trend analyses, correlation and regression analyses, estimation of variograms, normality checks. Transformations and/or filtering of the data that are retrieved from the database are also supported. The actual analyses are performed by the general statistical software program S+. This program is however not trivial to operate and does not work on the web. The SAT program can thus be regarded as an easy-to-use interface between the user on one hand and the capabilities of the S+ program, of which of course only a subset has been implemented and for which the input requirements have been specifically tailored to the contents of the database.

As a result, using the SAT program a first statistical analysis of the data is simple to execute. Furthermore the SAT program provides specific functions for the data manager to readily derive or update the parameters required in the generic tests used by SQC.

_ _ _ _ _ _ _

SPATIAL ANALYSIS

Usefulness of cartography

The three aspects of spatial analysis and cartographic tools and their usefulness for GIS consultation are presented and briefly discussed with regard to web request and processing possibilities. These aspects are the cartographic reference system, geographical objects properties and representation modes, and the relations between themselves. The three kinds of results - cartographic, graphical and tabular ones - returned as answer to the requests are then illustrated.

Pre-processed GIS data

Some physical parameters punctually measured during the campaigns are to be interpolated to produce continuous spatial information that characterises each of them. The interpolation processing method proposed is the universal kriging. This method requires the calculation of the semi-variogram and the adjustment of a specific curve for each case using the data collected at each station. The linear model is generally recommended. The interpolated parameter values are assessed in comparison with their computed variance values and on the basis of the continuously acquired data for some specific campaigns.

So the IDOD database can be completed by a set of pre-processed temperature and salinity grids in order to give to the users a quick look of the physical conditions.

A specific interpolation method has been applied to compute a 100m resolution Digital Elevation Model of the bathymetry in the Belgian continental shelf. The water volume conservation constraint has been defined, taken in consideration in this method and assessed (see poster presentation).

Other vector covers

Ancillary data must be stored in a joined database. For the time being, these data consist in coastlines. They were digitised at different accuracy levels and the merging process must take this fact in account. The other kinds of punctual, linear and polygonal features (wrecks, pipelines, biological areas, navigation routes...) could be inserted in this database.

Web applications

The majority of the data accesses will be performed through the network using a web browser. The user can access the data by filling a form to define the request criteria for selecting a subset from the IDOD database. This subset is returned to the user in a tabular text file format. Furthermore, it is possible for the user to get a cartographic representation of its results.

The other developed procedures provide to the users the possibility to perform some geo-processing applications like extraction of grid subsets, simple interpolation processes (Triangular Irregular Network, Inverse Distance Weighted methods), display and exploitation of data coming from them, etc.

For all these applications, we have chosen to use various products from ESRI. It requires programming of specific scripts adapted to the developed request forms.