Design of an oceanographic database

MULLER Fabrice[1], DE CAUWER Karien[2], SCHWIND Lucien[2], DEVOLDER Mia[2] & SCORY Serge[2].

[1]Laboratoire SURFACES, Département de Géomatique, Université de Liège, 14, place du 20 août, 4000 Liège, Belgique (Belgium). T : +32-4-366.57.45, F : +32-4-366.56.93.
Email : Fabrice.Muller@ulg.ac.be
[2]Management Unit of the North Sea Mathematical Models, 100, Gulledelle, 1200 Brussels, Belgium. T : +32-2-773.21.11, F : +32-2-770.69.72. Email : info@mumm.ac.be

1. The IDOD project

The Integrated and Dynamical Oceanographic Data Management (IDOD) project is carried out in the frame of the programme "Sustainable Management of the North Sea" funded by the Belgian Office for Scientific, Technical and Cultural affairs. Its purpose is to set up a national oceanographic information system, taking as basis the measurements made by the various projects funded by that programme. The IDOD project started in 1997 and its development and tuning phase will end in December 2001. Afterwards, the tools built in the course of the project will be kept within the Management Unit of the North Sea Mathematical Models, which plays the role of the Belgian National Oceanographic Data Centre.

1.1. Objectives

The purpose of IDOD is to establish, to manage and to promote a database of marine environmental data. Meanwhile, IDOD aims at ensuring a scientifically sound data flow between the data producers (laboratories executing routine monitoring, field and laboratory experiments, mathematical modellers and administrative authorities) and the end users (scientists, sea professionals, policy and decision makers, and the public in general). For a long time, various institutions and laboratories performed oceanographic measurements, often under co-ordinated research objectives. However archiving and management of the data was rarely co-ordinated. One of the objectives is to archive and make the existing data sets available in a homogeneous way.

The database will integrate different types of oceanographic data: data related to seawater, biota and sediment, continuous measurements (fixed position and underway), mathematical model results and geographical data. Quality control procedures are being applied on the incoming flow of data. In order to study the processes driving the marine phenomena, a set of data analysis tools is being developed. Various approaches are used: statistical techniques, geostatistics and spatial analysis. It is expected to take advantage of the information given by these analysis tools to improve the quality control on the incoming data.

Finally, one of the most important objectives of this project is to provide useful and scientifically sound information to a wide range of users. Special attention will be paid to derived products (maps, tables, reports, *etc.*) in order that they meet the specific requirements and level of expertise of the various categories of users.

1.2. Partners

Three organisations are involved in the project:

- Management Unit of the North Sea Mathematical Models (MUMM), a department of the Royal Belgian Institute for Natural Sciences;
- Laboratory SURFACES of the University of Liege (SURFACES);
- University Centre of Statistics of the Catholic University of Leuven (UCS).

SURFACES is responsible, together with MUMM, for the conception and physical implementation of the database and the development of the spatial analysis tools. UCS is responsible for the elaboration of the quality control and the statistical analysis tools in close collaboration with MUMM and the data providers. MUMM acquires and continues inventorying the different data sets by contacting involved laboratories.

2. Data description

2.1. Contact with data providers

Mutual confidence and contact between data providers and developers is of prime importance for the design of a database and data exchange. It is a long-lasting process. Several times a year, a meeting takes place with all co-ordinators of the different projects funded in the frame of the programme "Sustainable Management of the North Sea". Information is given on the progress of the IDOD-project and feedback is asked for. Most of the topics discussed during these meetings deal with theoretical aspects of the project, like the methods used to develop the database and the associated analysis tools. A newsletter is issued 2 to 3 times a year to inform all involved persons on the developments of the IDOD database. The papers published therein are send as "request for comments".

For practical information on the characteristics and the transfer of data, the IDOD-team takes contact with the data manager of each project. Project data managers are assigned for every project of the programme "Sustainable Management of the North Sea". Finally it must be noted that, in order to improve their knowledge of the data acquisition process, members of the IDOD-team regularly join scientists during sampling campaigns at sea.

2.2. Inventory and acquisition of data sets

The first task of the IDOD project consisted in inventorying the existence of relevant data sets. EDMED, the European Directory of Marine Environmental Datasets (which is currently being updated in the frame of *Sea Search*, an EU concerted action, [http://www.sea-search.net]), was used as a starting point for this task.

Although MUMM actively collects data at sea since years, its own data sets mainly concern contaminant concentrations in seawater. In order to have a complete description of the data types to be included in the database, it was thus necessary to first gather data sets pertaining to other parameters. It was decided to give priority to the data collected by laboratories involved in the programme "Sustainable Management of the North Sea".

Further investigation of these data sets had following advantages:

- they cover a wide range of disciplines;
- data providers can be contacted for more information on the characteristics of the data;
- missing meta-information can be asked for, as the projects are ongoing, resulting in an immediate impulse to laboratories for a better data management and quality control.

The following strategy is adopted: first of all, the involved laboratories are asked to submit recently collected data in the format considered appropriate by them. The IDOD-team examines every data set using a screening procedure. This procedure checks the availability of the data, the meta-information (date, time, position, sampling depth, methods and quality control information), expected parameters and the quality of the data set (clear description of parameter and unit, detection limit, precision, number of significant figures). During a visit to the laboratory, the missing meta-data and more explanations, especially on the methodology, is asked for. The laboratories receive feedback on the screening of their data.

The inventory and acquisition of data sets is a time consuming procedure. It took several months to receive from all services a list of parameters measured in the frame of the project. When finally a data set is received, it is mostly far from complete with respect to all required meta-information. The visits to the services proved to be very important. A lot of descriptive information is gathered and the necessity of meta-data can be motivated to the researchers during a personal contact.

3. Design methodology

3.1. An appropriate methodology

The design of a database for a GIS (Geographical Information System) requires a methodological approach of the problem. The most common scheme used to manage hybrid data sets is based on a couple of databases, one for the non-spatial alphanumerical data, and another one reserved for the spatial or geographical data. These two data sets constitute the data-side of the GIS. The design of a powerful GIS database must be elaborated with a reliable methodology. The MERISE methodology is one of the most advanced techniques to develop such a database [Nanci & Espinasse, 1996]. Therefore, in the framework of the IDOD project, it was decided to follow the design workflow suggested by MERISE.

Following the MERISE guideline, the conceptual datamodel of the database is realised using the Entity/Relationship formalism. This datamodel is a static representation of the information system or of all the data of the domain. It is designed with abstraction of the technical aspects of the database implementation and access techniques. It refers only to abstract objects and associations between these objects. The conceptual datamodel describes these objects and associations. There is no mention of what will be done with the data; only the semantic of the system is considered.

There are two ways to elaborate a conceptual datamodel:

1. a deductive approach that supposes the existence of a list of existing data to organise;
2. an inductive approach that aims to highlight all the concepts and ideas.

In the case of IDOD, the first approach is used because some archives of data are already existing.

3.2. Conceptual datamodel and E/R formalism

The Entity/Relationship formalism is a high-level semantic design tool that is set up on three fundamental constructive concepts: the *attributes* or *properties*, their gathering into *entities* and the links or *associations* between these entities.

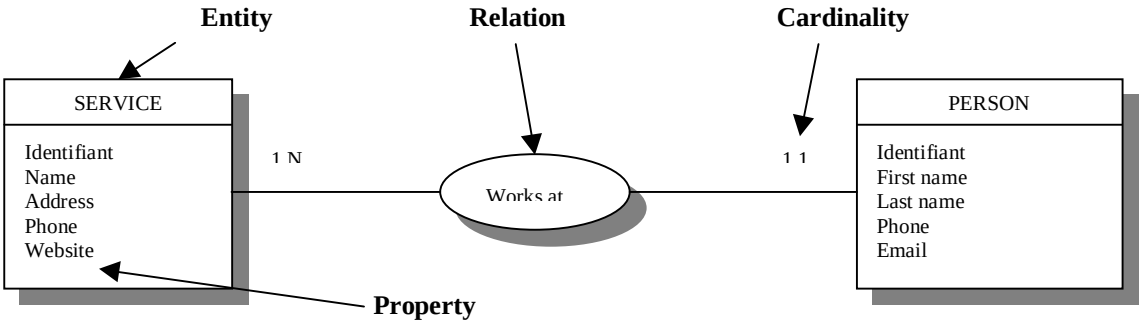


Figure 3.1. E/R formalism design.

Attribute or property

The *attribute* or *property* is the representation of an elementary or atomic information. It is the smallest part of information used by the system and having a meaning. The attribute can be single (first name, phone number, *etc*.) or compounded (address, date, *etc*.). The attribute describes the entity or the relation and exists only if it is part of an entity or a relation. An attribute is unique in a conceptual datamodel, so that it is part of only one entity or relation.

Entity

The entity represents a set of objects of the same type that are abstract or concrete. It is a group of attributes and represents only one semantic concept. The entity determines a type, a class or a group where all the elements are called the entity occurrences. To refer directly to one occurrence of an entity, the entity requires a special attribute called an *identifier* or *primary key*. One value of this identifier corresponds to one and only one occurrence of the entity, and the value of this identifier must be invariant and can never be changed or modified until the deletion of the occurrence of the entity.

Relation

The relation is a group of associations of the same type between two or more occurrences of entities of same or different types. The relation translates the words of the natural language. For example, in the figure below, the relation "*works at*" between the entities SERVICE and PERSON is the translation of the sentence *"person X works at service Y"*. At the opposite of

the entity, a relation has no real existence. The relation is simply expressed by the implied entities. The minimum number of entities involved in a relation is two. The dimension of a relation is the number of participating entities. Like an entity, a relation can also contain some attributes.

Cardinality

For each pair of entity-relation, the cardinalities are the minimum and maximum numbers of occurrences of the relation that can exist for one occurrence of the corresponding entity.

The cardinality values are indicated upper or on the link between the entity and the relation. The most common values for cardinality are:

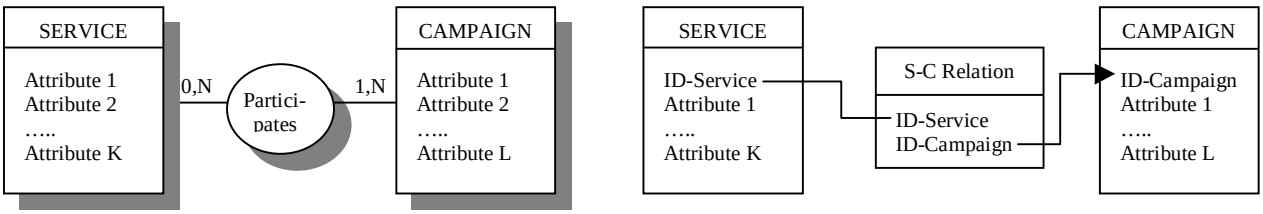
- **0,1**: one occurrence of the entity can exist without participating in the relation (0), and if it participates it is only once (1);
- **0,N**: one occurrence of the entity can exist without participating in the relation (0), or can participate without limitation (N);
- **1,1**: one occurrence of the entity participates one and only one time in the relation;
- **1,N**: one occurrence of the entity participates at least once in the relation.

Many other extensions exist in this formalism such as the constraints, specialisation, generalisation, *etc*. but they will not be presented in this paper.

In practice, the demonstration version of the case tool *Dbmain* [Hick *et al*., 1999] is used for the design of the conceptual datamodel. The conceptual datamodel, presented in annex 1, is produced with this software.

3.3. Logical datamodel

The logical level is the second step in the database design. The entities are converted into tables for a relational database. There are three kinds of relations to be considered: one to one, one to many, many to many. The first and second types of relations are converted into the logical datamodel by adding new foreign keys to the tables. The third one is more complex and requires the creation of a transition table as illustrated in the figure 3.2. A service may participate in several campaigns, but a campaign involves at least one or more service(s).



Conceptual datamodel. Logical datamodel.

Figure 3.2. Example of relation.

4. Design of the IDOD database prototype

4.1. Semantics of data

To organise the data, the different steps from collection till result have been reviewed The resulting structure is generalised in figure 4.1, the detailed model is shown in Annex 1. 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. A project results into or contributes to one or more data sets and vice versa. Samples are characterised by a position, sampling depth, time and sampling method. The monitoring year, the sequence number or sampling occasion number and a sample code identify a sample. Information on meteorology and the ecosystem (*e.g.* estuary, coastal sea) will be stored with every sample to enable the direct display of the environmental circumstances.

The resulting values are stored in the original units to maintain the number of significant figures. The values are accompanied by a qualifier flag indicating if the reported value should be qualified as "less than" or "greater than", a validity flag given by the data provider or manager, and a statistical quality control flag. In order to obtain the latter, the consistency of values is checked against the statistical characteristics (like spatial continuity of data and the correlation between variables) of the results already present in the IDOD database. 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. Both flags are especially necessary for historical data.

The analysis method and sample handling method (preservation, separation, pre-treatment) are related to the result. For the description of the analysis method, character codes will be appended in one string. This way the different characteristics of importance for a certain method can be reflected in a flexible way. More information can be given in the description. Detection limit, detection unit, and precision are foreseen in the entity "analysis method". Every value is linked with a parameter described by a code and a full name, accompanied by the measurement unit, matrix and substrate. The parameters are divided in parameter groups.

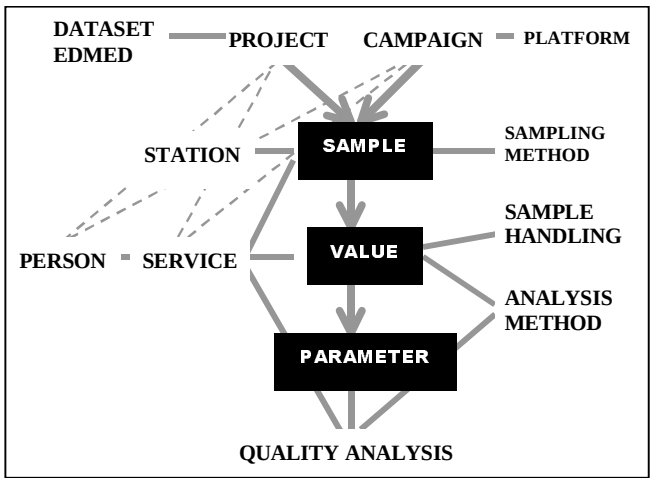


Figure 4.1. General structure of IDOD for data measured in seawater.

To incorporate information on the quality of the analysis method, three different aspects have to be introduced: QUASIMEME intercomparison exercise codes (Topping, 1998), control chart information and intercalibration exercises. For a certain service measuring a certain parameter with a certain analysis method, one or more scores can be obtained during intercomparison exercise(s), and similarly for information on control chart(s) and intercalibration exercises.

4.2. Conceptual datamodel

Following the semantics of data, a conceptual datamodel was elaborated for seawater. It is shown in Annex 1. The *DB-MAIN* case tool was used for the visualisation of the model. Twenty-one entities and twenty-nine relations compose this model.

Some entities are linked together by more than one relation such as "project" with "person" because there is one co-ordinator for a project and many other persons working on this project. The cardinalities are used to express these constraints (*e.g.* one and only one person is co-ordinator for a project, but a person can supervise from zero to N projects). To avoid mistakes and inconsistency, the three first normal forms of the database conception rules were checked and some modifications were applied to satisfy them [Kroenke & Dolan, 1988]. With respect to the third normal form, the "parameter" entity was split into two new entities: "parameter" and "category". Many parameters can be in the same category, and this way, it is possible to modify a category without having to browse all the parameters included in this category to apply the change. It insures the integrity and consistency of the database.

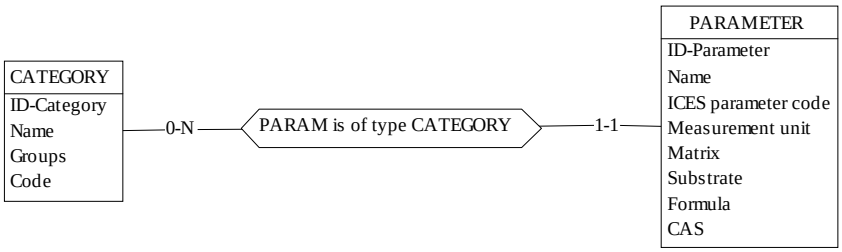


Figure 4.2. Example of splitting one entity into two entities.

As mentioned above, there are three types of information related to the quality of analysis to be incorporated: QUASIMEME intercomparison exercise codes, control chart information and intercalibration exercises. All these are represented in the model by use of the specialisation with the constraint "T" specifying that an occurrence of the entity is specialised into one and only one sub-entity.

A quite similar datamodel was elaborated for continuous measurements (like automatic cruise profiling). The major change is the removal of the entity "sample" which has no sense for these values. There is a direct relation between the entities "campaign" and "continuous value".

4.3. Data dictionary

The conceptual datamodel is closely linked to the data dictionary. The dictionary is the user's guide of the database with a complete description for all attributes and entities. A short example for the entity "station" is shown in figure 4.3.

DATABASE: IDOD				**ENTITY : STATION**	
DESCRIPTION : The STATION entity describes a geographical location sampled regularly.					
RELATED ENTITIES : NON-CONTINUOUS VALUE, CONTINUOUS VALUE.					
ATTRIBUTES :					
NAME :	**EXPLANATION**	**TYPE**	**FORMAT**	**EXAMPLE**	**COMMENT**
ID-Station	A code to identify one occurrence of the STATION entity in the database.	Key			Required, unique, system-generated
Name	The name of the station.	String	10	330a	Required
Start date	Date of first sampling	Date		21/11/85	
End date	Date of last sampling	Date		07/12/95	
Reference latitude	The latitude of the station location.	Float		51.43333°	Required Decimal degrees
Reference longitude	The longitude of the station location.	Float		2.80833°	Required Decimal degrees

Figure 4.3. Description of the entity STATION in the data dictionary of IDOD.

For each attribute, the dictionary mentions the field type and format, and presents an example accompanied by a comment. When "Required" is mentioned in the "Comment" column, it means that a value is mandatory for this field.

4.4. Logical datamodel

The logical datamodel is obtained from the conceptual datamodel here above. The entities are converted to tables and some conversion rules are applied to the relations. As explained in section 3.3, it required the adjunction of new attributes or tables to materialise some types of relations. For the prototype, all these steps are made manually. With a case tool, they would be generated automatically.

4.5. Physical implementation

First, a prototype of the IDOD database was developed. For the prototype, the choice of the DBMS *Microsoft Access97* was motivated by a friendly interface to generate the tables, forms, requests, states and macros. With its ODBC (Open Database Connectivity), it also offers the connection and sharing of data with many other software's. Moreover, this DBMS was available for every partner of the project for a low price. The migration to the final DBMS (*Oracle, Sybase, SQL Server, etc.*) will be done easily by use of existing exportation

filters. The tables and links are implemented in respect with the previously sketched logical datamodel and the cardinalities are implemented in *Access97* by the property values attached to each field of the tables.

A first batch of data has been entered in the prototype making use of export tables of the previous database (MONITB) available at MUMM. Besides this, extra information had to be entered like the different projects, campaigns, persons, *etc*. not stored in MONITB. Especially this last process is laborious. In future, ROSCOPS forms will be used for the input of information and for validation of MONITB data.

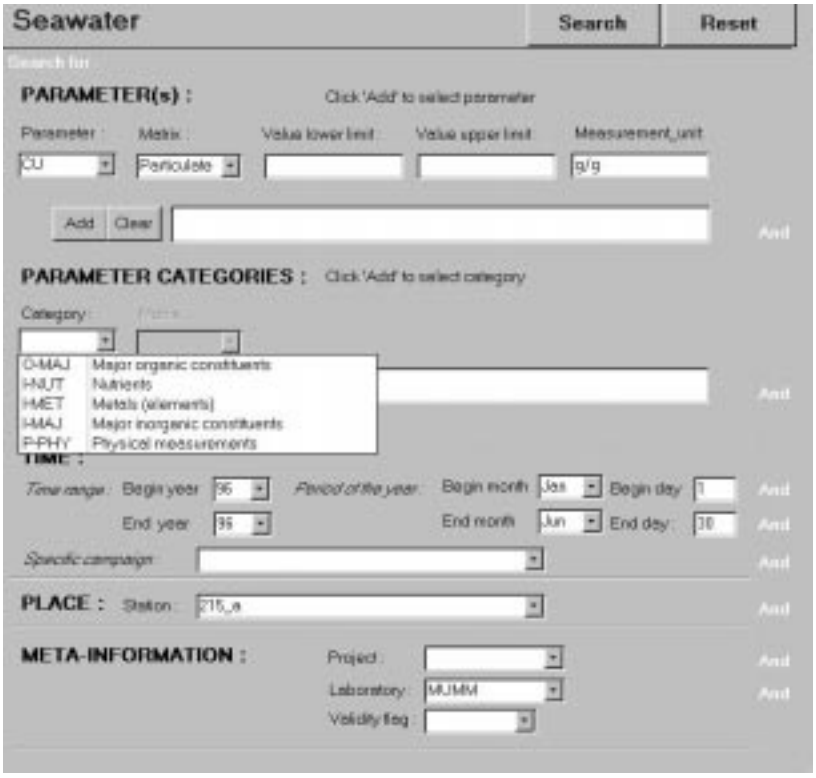


Figure 4.4. Form to select seawater data.

A general query has been constructed within an Access form as illustrated in figure 4.4. *Microsoft Visual Basic* was used to construct the SQL-code [Viescas, 1997].

5. Discussion and conclusion

The IDOD-project aims at integrating a high diversity of oceanographic data. A good understanding of the data is a prerequisite for the design of a database. Therefore, a regular contact and several discussions with the data providers were planned. Visits to the data providing laboratories resulted in a lot of descriptive information, while at the same time the need for meta-information could be motivated.

Acquiring data and all meta-information proved to be a very time-consuming process. Data providers have to be contacted several times before a data set can be considered as complete.

Consequently, it was decided to start with the development of a prototype for the concentration data measured in seawater, as a lot of information was already available at MUMM. In the resulting prototype, the values are accompanied by documentation enabling the user to assess the quality and suitability for a particular task.

6. Perspectives

Similar datamodels are being developed for sediments and biological data. Continuous measurements will also be included in the database. The spatial or geographical data will be organised in a GIS (Geographical Information System) such as *Intergraph Geomedia* or *ArcView*. These data will be recorded in the internal format of the GIS software.
As they offer two important characteristics, namely their capability of simulating and anticipating processes and events, mathematical models are thought to be also an important source of information in the scope of the project. Adequate validation procedures will be defined and results of relevant and validated models will be incorporated into the database.

To improve the quality and rapidity of the database design, it is planned to acquire a case tool offering a complete workflow from the conceptual model to the physical implementation into a DBMS. A benchmark was achieved to compare about ten case tools available on the market. A few of them (*WinDesign*, *EasyCase*, *Silverrun*, *AMC Designor*, *etc*.) emerged and are currently being further investigated.

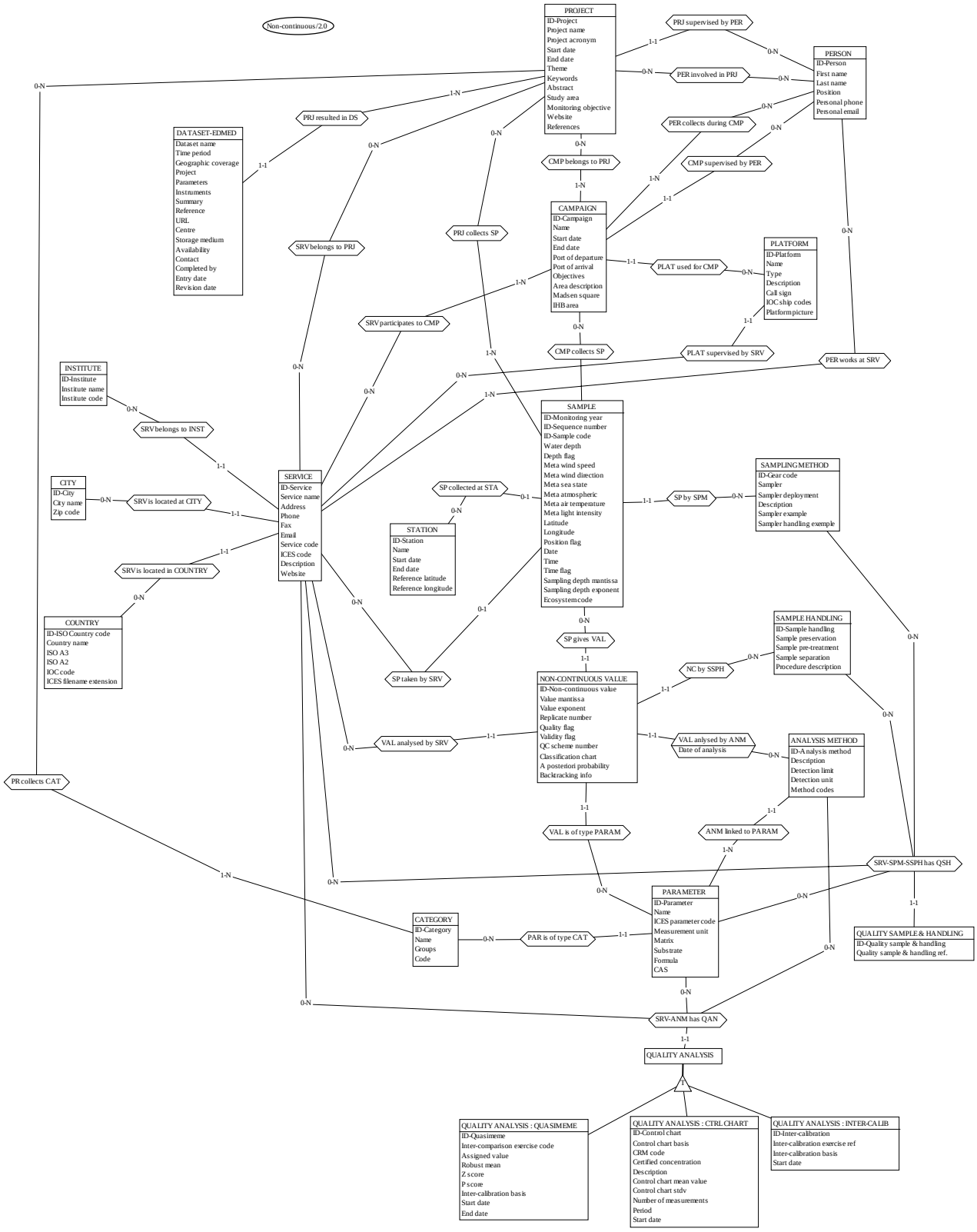
7. Acknowledgements

The IDOD project is fully sponsored by the Belgian federal Office for Scientific, Technical and Cultural affairs.

8. Bibliography

- Hick J.-M., Hainaut J.-L., Englebert V., Roland D., Henrard, 1999. *Strategies pour l'evolution des applications de bases de donnees relationnelles : l'approche DB-MAIN*. XVIIe congres Inforsid, La Garde, France, June 1999.
- Kroenke D., Dolan K., 1988. *Database processing : fundamentals, design, implementation*. Science Research Associates, 687 p.
- Nanci D., Espinasse B., 1996. *Ingénierie des systèmes d'information Merise*. Sybex, 891 p.
- Topping G., 1998. *Introduction to the Special Issue of Marine Pollution Bulletin*. Marine Pollution Bulletin, Vol. 35, Nos 1-6, pp. 1-2.
- Viescas J., 1997. *Microsoft Access au quotidien*. Microsoft Press, 969 p.

- Annex 1



Conceptual datamodel for the seawater.