Mediterranean Marine Science
Indexed in WoS (Web of Science, ISI Thomson) and SCOPUS
The journal is available on line at http://www.medit-mar-sc.net

Paper presented at the
"International Conference on Marine Data and Information Systems - IMDIS", Athens, 2008

OGS NODC Meta-data standardization, discovery and reporting from a relational database

M. VINCI, A. GIORGETTI and A. BROSICH

Istituto Nazionale di Oceanografia e di Geofisica Sperimentale,
Borgo Grotta Gigante 42/C, 34010, Sgonico (TS), Italy

Corresponding author: mvinci@ogs.trieste.it

Abstract

The use of technology constitutes a critical aspect in the oceanographic data management like in other research fields. The Department of Oceanography of the OGS has a high level experience in collecting oceanographic data obtained through involvement in several national and international programmes. The Oceanographic database managed by the OGS National Oceanographic Data Centre archives measurements of physical and biogeochemical parameters, of current and wave motion, of sea levels and meteorological data described by correlated meta-data from the beginning of the 1900's until now. All the data and meta-data are currently contained in an Oracle relational database. Access to all the meta-data is public whereas access to the data is subject to a Data Policy defined at data set level. In accordance with the SeaDataNet project, whose target is to create an integrated network of European oceanographic data centres, the OGS NODC is developing a standardization of the meta-data using XML and Web Services technologies thanks to the high flexibility of its relational database.The OGS NODC provides a discovery and reporting service through dedicated web pages that allows different kinds of users to obtain information about the oceanographic field they are interested in.

Keywords: Meta-data; Oceanography; Standardization; Interoperability; Relational Database; Marine Data; XML; XSL; Web Services.

Introduction

The Department of Oceanography of the OGS has a long-standing and high-level experience in the collection of oceanographic data obtained through participation in several national and international experimental and applied programmes.

The development of an efficient marine data bank has been a necessary step for the administration of a wide range of data. In 1979, a first relational database model was implemented by the OGS Informatics Centre to manage the meta-data obtained from the ASCOP programme (MANCA, 1979). The beginning of this decade saw the introduction of the Oracle relational database, thanks to its flexibility, performance and widespread use.

Since June 2002, the OGS has been the National Oceanographic Data Centre (NODC) for Italy (MOSETTI, 2003), giving it an institutional role in the Italian data management. (GIORGETTI, 2007).

The Oceanographic relational database managed by the OGS National Oceanographic Data Centre archives measurements of physical and biogeochemical parameters, of current and wave motion, of sea levels and meteorological data described by correlated meta-data. This information is validated and continuously updated in the relational database. Access to the meta-data is public whereas access to the data is subject to a Data Policy defined at data set level, in agreement with the data providers' guidelines.

The oceanographic meta-data archived in the database include information related to:

CSR - Cruise Summary Reports

The Oceanographic Cruises Inventory maintained at OGS gathers more than 150 Cruise Summary Reports, the oldest cruise dating back to 1909. Most of the cruises were carried out on board Italian Research Vessels. Some reports were also obtained from cruises involving joint scientific projects or ship exchange programmes with foreign organisations. The inventory aims at facilitating the search for information and data collected during the research cruises. Each cruise is briefly described by the chief scientist(s) with regard to objectives, work carried out, geographical area, location track, bibliographical references. The works cover several disciplines: marine geology, geophysics, physics, chemistry and biology.

EDMED – European Directory of Marine Data Base

The Marine Data Inventory describes more than 500 Italian marine data sets or international databases of general interest, collected by several Italian scientific laboratories. The inventory covers physical oceanography, chemical oceanography, biological oceanography, marine meteorology, hydrography, marine ecology and underwater acoustics.

EDMERP - European Directory of Marine Environmental Research Projects

The Research Projects Inventory describes Project Reports relating to the marine environment. The inventory covers a wide range of disciplines.

Initially imported from other databases (Access) and storing devices used in the past, using Oracle features, now this information is included for a better exchange of meta-data between European Data Centres.

The OGS NODC relational database and Technical details

Since 1970, thanks to the studies of Dr. E.F Codd ("*A relational Model of Data for Large Shared Data Banks*"), the structures of the database and correlated database management systems (DBMS) was changed to a more efficient and flexible ,not hierarchical, model (CODD, 1970).

A relational model is an alternative to the old hierarchical structures. Like other database models, the relational one is based on *records.* A record is considered to be an elementary set of data which lets one define an entity. These records are correlated between each other inside *tables.* A table is composed of rows *(tuples)* that correspond to the records and columns that are like the fields of records. In a relational database, *a* subset of records is descriptive for the structure, for this reason the database is called self-descriptive (information about data relations, constraints,etc...). This lets us access all the applications able to read the information contained in these self-descriptive-records. A data structure as described above, has automatic mechanisms inside it able to manage records, and for this reason is called *Relational Database Management System* RDBMS.

An RDBMS allows you to:

- add one or more new tables in a simple way correlating them without modifying the database structure and without modifying the database management applications.
- Lets one modify the structures of tables in a very simple way.
- All the RDBMS are accessible using the same tool, the *SQL* language available for the main number of hardware platforms.

The Oceanographic database of the OGS

NODC is implemented using an Oracle 11g Database on a Debian GNU/Linux installation. It contains more than 60 tables; 16 of which are needed to save the measurements and the metadata included in MEDAR/MEDATLAS data format, while the others have been added to address the needs coming from the SeaDataNet standardization. (Oracle).

The database now archives more than 120 million in-situ measurements, using almost 75 GByte of disk space also used for data-warehousing needs. Due to the large amount of measurements archived, some kinds of aggregations of data are required to achieve a good standard of performance for the querying data. These aggregations are obtained using dynamic SQL within stored procedures or materialized views periodically refreshed.

Standardization

According to the SeaDataNet project, whose target is to create an integrated network of European oceanographic data centres, the OGS NODC is developing a standardization of the information acquired to date. A first step in this direction is to obtain an high Interoperability between systems used by partners and to develop a common format and values for the meta-data. (SeaDataNet).

We can simply say that Interoperability is the ability of heterogeneous systems and organizations to work together. The interoperability of software is a term used to describe the capability of different programmes to exchange data via a common set of exchange formats, to read and write the same file formats and to use the same protocols (Wikipedia).

The standardization is based on the conversion of the information contained inside the database to an XML standard. Inside these XML documents, information obtained from centralized standard vocabularies, accessed through the Web Services technology, is included. This is done to exchange and harmonize the information with the other partners.

XML stands for EXtensible Markup Language, which is a markup language much like HTML. It was designed to transport and store data, not to display them. Its tags are not predefined and you must define your own tags (W3C).

Standardization through XML follows the standard ISO ISO19115 reached by the OGS NODC using the XSL language.

XSL stands for EXtensible Stylesheet Language and describes how the XML document should be displayed (W3C). The World Wide Web Consortium (W3C) started to develop XSL because there was a need for an XML-based Stylesheet Language. Two of the XSL language tools where important to transform the data contained in our database into the XML ISO ISO19115:

- XSLT - a language for transforming XML documents
- XPath - a language for navigating in XML documents

The information contained in the database is extracted producing a first XML tree shape and then transformed using the appropriate XSLT to obtain the desired output.

More than one XSL stylesheet follows the first output obtained from the database extraction. Using the XPath it is easy to find the different branches of the XML tree to give them the final shape.

Standardization through the web services lets one use a common language between the partners involved in the European project and reach the target of a higher automation on information exchanges. One of the final targets will be to let Web Services of different data centres interact between each other.

A Web Service is a software system designed to support interoperable machine-to-machine interaction over a network. Software applications written in various programming languages and running on various platforms can use Web Services to exchange data over the Internet in a manner similar to inter-processing communication on a single computer (W3C).

To deploy Web Services in the SeaDataNet on different platforms, a common Web Service architecture was defined. Some characteristics

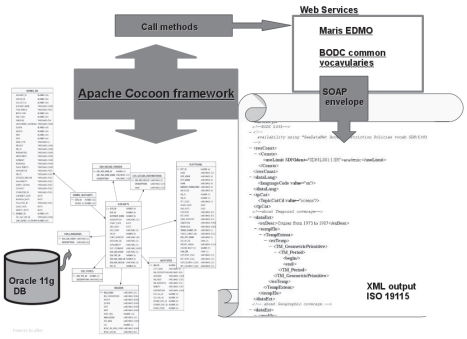

Fig. 1: Meta-data standardization through Web Service direct call.

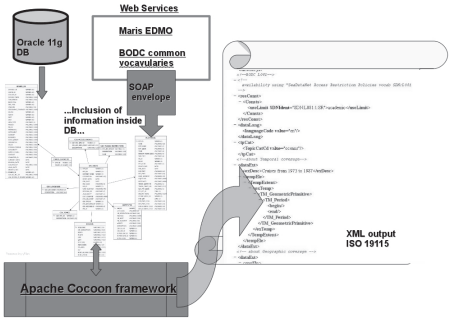

Fig. 2: Meta-data standardization including Web Service inside database.

are the use of HTTP protocol and a standard XML data format.

The Web Services contribute to developing the standardization through two main sources:

- Maris: EDMO Web Services that contain information about "marine organizations",
- BODC: that manages the common vocabularies that contain all the standard terms about parameters (P021), ships (C174), data policy (L081), etc.

Some predefined "call methods" let the users query the Web Services with the appropriate standard code included in the database and obtain information about: standard terms, versions and last modification date of the vocabulary.

Two architectural styles were possible for their implementation :

- RESTful services defined as an ROA (Resource Oriented Architectures) which connects a unique URL to a Resource.
- SOAP-based services defined as an SOA (Service Oriented Architectures) which interact with predefined "call methods".(Whatlist (), GetList (),etc.)

We adopted SOAP because its rigid, formal structure allows a clear path for the software development also thanks to the use of a widely-adopted Web Service toolkit.

Two ways were followed to include this standard information in the XML document.

The Web Service was queried directly for each term required for the EDMED standardization (Fig. 1). The inclusion of these results in the XML stream was done using a (Cocoon) feature (<cinclude>).

Inclusion of the data, downloaded from Web Services inside the database with a stand-alone software for their update was followed for the CSR and EDMERP standardization (Fig. 2). In this way, the information was included in the output document using a simpler and faster SQL query.

A validation tool (http://www.seadatanet.org/validator/) developed inside the SeaDataNet project allows one to check the accuracy of the XML schema obtained and the terms used. The XML files are loaded from the local file system through the validator web page and checked using an updated XSD schema. This lets the user obtain information about the accuracy of the XML schema and the uniformity of the standard terms used. In addition to the Schematron the OGS NODC uses a commercial software (Oxygen 9.3) to make a crossed validation. (OXYGEN).

Discovery and Reporting

The relational structure of the OGS NODC database lets one generate a wide range of output aggregating information from many tables (data-warehousing) through appropriate *SQL*

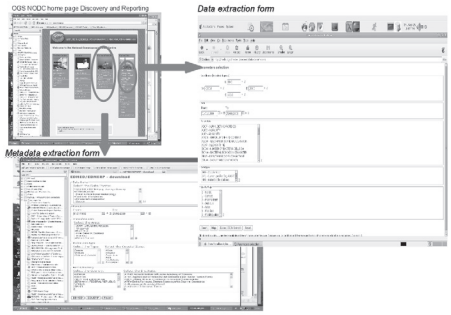

Fig. 3: The OGS NODC home page with Discovery and Report forms.

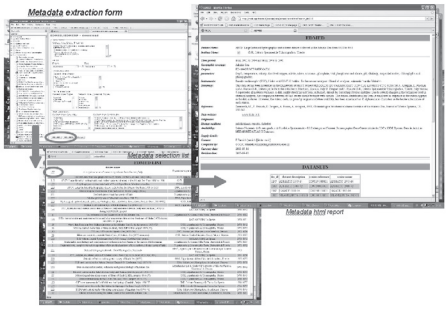

Fig. 4: Meta-data Discovery and Reporting service.

query. Some "materialized views", a kind of table that can be easily updated and, that aggregates data of time series (current meters) or data of vertical profiles (CTD), were generated for the output of data. This is a first step to querying very huge tables that contain hundreds of millions of records in an efficient way.

For the output of the meta-data, some nested "select query", with the appropriate links (join) between tables, are able to generate the necessary aggregation.

A combination of SQL query generated dynamically using some cocoon tools and XSL stylesheets,can produce a Web output. The information is extracted from the relational database (using SQL query via JDBC connection). An XML framework like Apache Cocoon is used to generate, to combine and to transform everything into a single output document with the requested format using a variable number of XSLT transformations.

Thanks to the previous features, OGS NODC developed a portal that supplies dedicated web pages for different meta-data European Directories (http://nodc.ogs.trieste.it/nodc/homepage). This lets the users search data and meta-data by prefixed research criteria like time period, institute, data theme or data type. The result of the search is a list of records which satisfies the previously selected criteria. These web pages together with the web pages dedicated to data research can reach a wide range of users

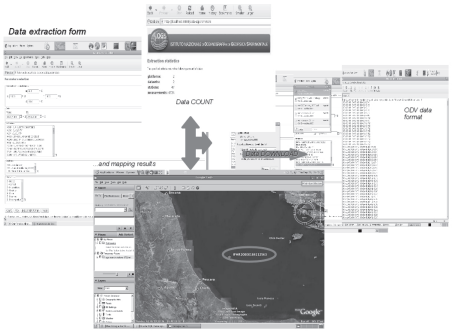

Fig. 5: Discovering and mapping the data.

interested on the oceanographic field (Fig. 3)

Currently three web pages for the search of meta-data about European directories are available (Fig. 4):

- CSR
- EDMED:
- EDMERP

Two other forms where created for data searches (Fig. 5):

- data sets
- parameters

Conclusions

After a long-standing experience as data collector, the assignment to National Oceanographic Data Centre gave the OGS Oceanography Department a central role in the management of Italian oceanographic data. The institute is mak-

ing an important contribution at European level by participating in the marine-data interoperability projects.

The core of the technical management in OGS is represented by the use of an Oracle relational database. With its non-hierarchical structure one can now store, update and efficiently manage a huge amount of information. New kinds of information can be added without greatly altering the existing structure. For the near future, a more efficient and user-friendly database updating tool will be implemented with the help of software based on widely adopted java technologies.

The flexibility of the Oracle database on data extraction,in addition to features of the EXtensible Stylesheet Language lets one develop web applications like discovery and report web pages. These allows users to obtain information about data and meta-data following the research criteria suggested.

To improve interaction with users the next step will be to optimise the querying tools and map the results of the data reporting with a graphical output. This will be done with the help of OLAP (On Line Analytical Processing) tools and Geographic Information Systems or interpolation software (DIVA).[DIVA]

Procedures to migrate reaggregated data (data-warehousing) to RDBMS with spatial extensions will support an easier data analysis.

Acknowledgements

This work is part of the project EU-SeaDataNet contract n.RII026212. Part of this work was presented in a poster format at the "IMDIS" meeting of April 2008 in Athens.

References

GIORGETTI, A., BROSICH, A. & MOSETTI, R., 2007: *OGS oceanographic data archiving and validation system: the IOC/National Oceanographic Data centre.* Bollettino di Geofisica Teorica ed Applicata Vol. 48.

COCOON: http://cocoon.apache.org/

DIVA: http://modb.oce.ulg.ac.be/backup/modb/diva.html/

Dr. E.F Codd, 1970: *A relational Model of Data for Large Shared Data Banks.* Communication of ACM

NODC: http://nodc.ogs.trieste.it/nodc/homepage

ORACLE: http://www.oracle.com/index.html/

OXYGEN: http://www.oxygenxml.com/

SEADATANET: http://www.seadatanet.org/

W3C: http://www.w3c.org/

WIKIPEDIA: http://wikipedia.org/