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

Multi-standards metadata cataloguing tools for Ocean Observatories

J. BARDE, J.-C.DESCONNETS, D. EGDINGTON and V. HEURTEAUX

IRD, Centre de Recherche Halieutique Méditerranéenne et Tropicale (CRHMT),
Avenue J. Monnet, BP 171, 34203 Sète, France
MBARI, 7700 Sandholdt Road, Moss Landing, CA, 95039-USA
Géomatys, 24 rue Pierre Renaudel 13200 Arles, France

Corresponding author: julien.barde@ird.fr

Abstract

Nowadays, information systems interoperability is made possible by using Internet and related standards (with XML schemas) to share metadata, data and services. However, most of the time, the users complain about issues to get relevant informational resources (IR) for their work. Indeed, most of them are still distributed into heterogeneous systems whose interoperability (IR discovery, access and treatment) faces a lack of standardization. This situation is critical in domains where fundings and people available for IR management are limited. Moreover, it is obvious that none of existing standards can cover all these needs: according to the various domains and related kinds of IR, different standards are required (one per domain or kind of IR in the best cases). Furthermore, some issues remain whatever the standards to implement, like describing semantic or spatial content of IR in a machine understandable way. The challenge thus consists in setting up kinds of information systems built on top of existing ones which enable a mutli-standard management (and thus interoperability with partners) and limit additional development efforts. By using a generic approach, we suggest models to manage different kinds of standardized metadata described by common controlled vocabularies (for thematic and spatial descriptions) within a single architecture. We aim to set up such a tool in the case of ocean observatories to manage different kinds of metadata elements sets (previous and new reference standards: SensorML...).

Keywords: Information systems; Metadata; Ontologies; Interoperability; Sensors.

Introduction

This paper presents the goals and first results of an ongoing data management project at MBARI (Monterey Bay Aquarium Research Institute) which aims first to facilitate Informational Resources[1] (IR) discovery by improving its metadata management. According to institutes and projects, IR are produced in heterogeneous ways but are needed together (as inputs for different kinds of studies related to the marine

1 Any kind of data, information, knowledge produced (including the treatments) to study a domain (regardless of the format: numeric or not). According to the domain and users: an observation, a report, a map, a picture, a video, a dataset/serie, a database, a model, a data measured by a sensor, an interview, a knowledge...

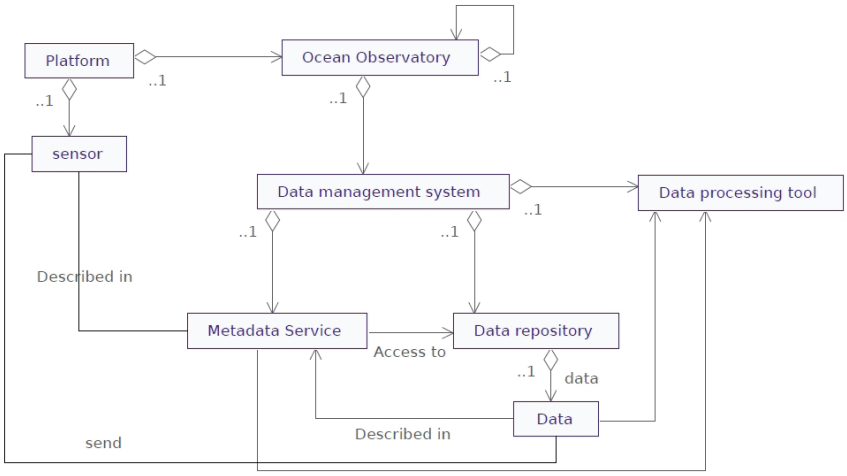

Fig. 1: Ocean Observatories as a UML classes diagram.

domain). Currently, in order to easily describe and retrieve IR, users should implement reference metadata standards. These standards aim to support the interoperability with similar systems located around the world (usually by using Web Services) and thus facilitate data harvesting in a more efficient (exhaustive) way. Moreover, since management of spatial dimension is a crucial component of IR in the environmental domain, the relevant metadata standards have to deal with the ability to manage spatial information. These standards are mainly related to the common work of ISO TC/211 [ISO] and OGC [OGC] (and compliant with W3C recommendations). These issues are well known in the case of environmental domain, especially in the marine domain (KAZAKOS *et al.*, 2002) and in the set up of Ocean Observatories: there is a strong need to improve IR sharing for users to work properly. Nevertheless, as theseOGC standards are recent as well as complex, institutes are still using either previous standards (which are rapidly becoming obsolete like FGDC whose scope is similar to the new OGC/ISO 19139 standard but can't handle interoperability at a global scale) or more specific standards (like Thredds metadata element set [CARON, 2004] to describe the content of files managed with the Netcdf data format and shared by using OpenDap protocol) or "home made" metadata elements sets created to cover local needs. Most of the time these different metadata elements sets are managed within different information systems based on different architectures (whose structures are specifically adapted to the content of the standard and related profiles they aim to manage) (BERNSTEIN *et al.*, 2000). Therefore, we propose a generic approach to manage any metadata standard into a single architecture built on top of existing information systems. In the section "Issues of information management", we summarize the main issues to set up an efficient IR management and some specific needs of Ocean Observatories. In "Generic models for metadata and semantic management", we suggest to implement generic models to manage different kinds of metadata standards as well as controlled vocabularies. In sections "Ongoing implementation" and "Homogeneous GUIs set", we present examples of such models implementations (ongoing developments) and some related GUIs which demonstrate the interest of this approach from a user point of view.

Issues of information management

Since 20 years MBARI set up different ocean observatories by using different technolo-

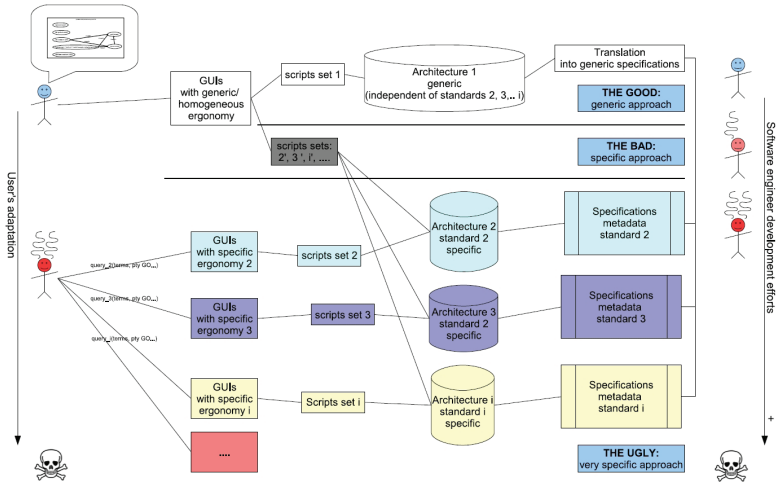

Fig. 2: Reducing developments efforts by using a generic approach (Barde et al., 2007).

gies for softwares and hardwares (platforms, sensors...) (GOMES *et al.*, 2006, O'REILLY *et al.*, 2006). 2.1 Ocean Observatories and information management Ocean Observatories are made of different kinds of platforms which are themselves made of different kinds of components (sensors...) and are deployed in different locations (Fig.1). According to the period and projects, these observatories have used the most relevant technologies (among available ones) to acquire and manage sensors data into different information systems: different metadata, different formats and implementations (Fig.1). MBARI is currently managing at least 4 different kinds of metadata elements sets (ISO 19115, Thredds, SSDS and recently SensorML) to describe IR and their acquisition hardwares (platforms, sensors...). These metadata are managed with different formats. An efficient data discovery is currently complicated...and thus data access is obviously difficult, especially for external users.

Instead of modifying each (previous or current) information system, the challenge consists in building new kinds of information system on top of them which enables to import and manage any kind of metadata elements set (including previous ones). This way users will be able to retrieve and access informational resources (managed in different information systems) by handling different kinds of metadata through a single portal. While ISO/TC 211 and OCG set up reference standards for spatial information (for example ISO 19139, SensorML for metadata, O&M, GML, KML for data, CSW, WMS, WFS, WCS, SOS for Web Services) interoperability, we aim to make current information systems compliant with them by using a generic approach.

The good, the bad and the ugly

Figure 2 illustrates some of the issues due to heterogeneous implementations of (similar or different) metadata standards. Usually a new information system is set up for any new implementation of a metadata standard (with new databases physical models, scripts sets and GUIs). This is the reason why software (engineers) need to minimize their developments efforts by using a generic approach to satisfy the same user needs in the same way.

Whatever the standard implemented, we aim to manage it by using a single scripts set and the same components. The related architecture has to implement a generic architecture to reach this goal (BERNSTEIN *et al.*, 2000). 2.3 Different metadata standards with a similar core

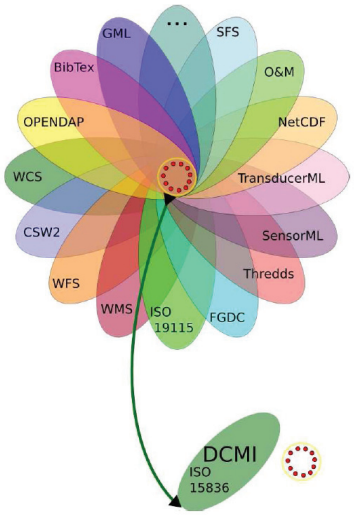

Fig. 3: Different metadata standards with similar core metadata elements (BARDE *et al.,* 2007).

According to IR that users need to describe, standards are made of different metadata elements. However, these different standards share some core common metadata elements which match the Dublin Core (DCMI) metadata elements set (including spatial description, Fig. 3) which answer the basic questions in Information Research (Where? What? When?...). Nevertheless, there is no single generic metadata standard enabling any kind of IR description. As environmental studies require different kinds of IR, they need then to implement different standards. There is thus a lack of tool to import and manage different standards within a single application.

Generic models for metadata and semantic management

This section presents the generic models we suggest to manage different meta data standards and controlled vocabularies into a single architecture.

Conceptual model for multistandard metadata management

By using a generic pattern (conceptual model) we aim to describe in the same way any metadata standard. We can then use this model to set up a multistandard metadata management tool assisting[2] the valuation of {metadata element, value} pairs.

The UML class diagram in Figure 4 illustrates our approach. Any metadata standard can be considered as a collection of metadata elements related by different kinds of relationships (generalization, association...). Most of the time, a standard can be adapted (profile) by removing (or expanding) metadata elements in the native set. Each element can get a free text or controlled value. Nevertheless, core metadata elements have to be controlled to improve IR management.

Additional model for semantic management

Whatever the standard, heterogeneous values of core metadata elements in-terferes with IS semantic interoperability. In particular, the use of common semantic and geographic/spatial referential is a key issue to control thematic and spatial descriptions. This is useful at a local scale but required at a global scale to share the meanings of descriptions with external users:

- the management of "keywords" like metadata elements requires a tool to control the terminology of concepts used for their values. A glossary, the- saurus or ontology is needed to improve data discovery. Indeed, it is impos- sible to match the content of different metadata sheets without managing multilinguism, synonyms... Figure 5 expands the previous UML di- agram by incorporating a set of classes (Semantic management package) which enables the management of structured concepts schemas (ontologies and thus thesauri or glossaries) (BARDE *et al.*, 2006),
- the spatial description package of Figure 5

2 not directly into XML files.

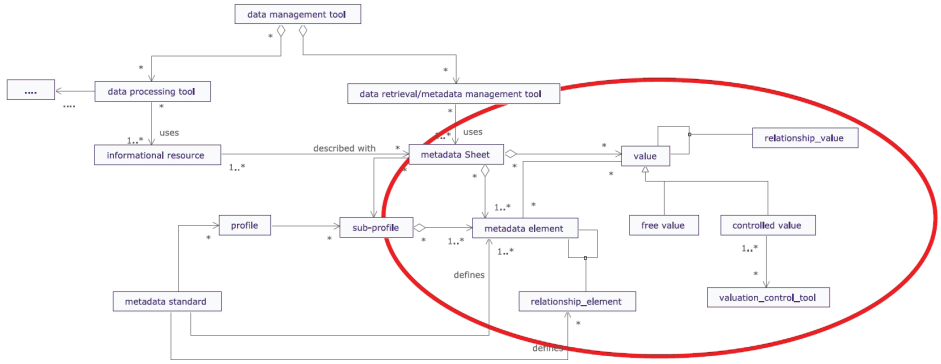

Fig. 4: Generic model for multi-standard management.

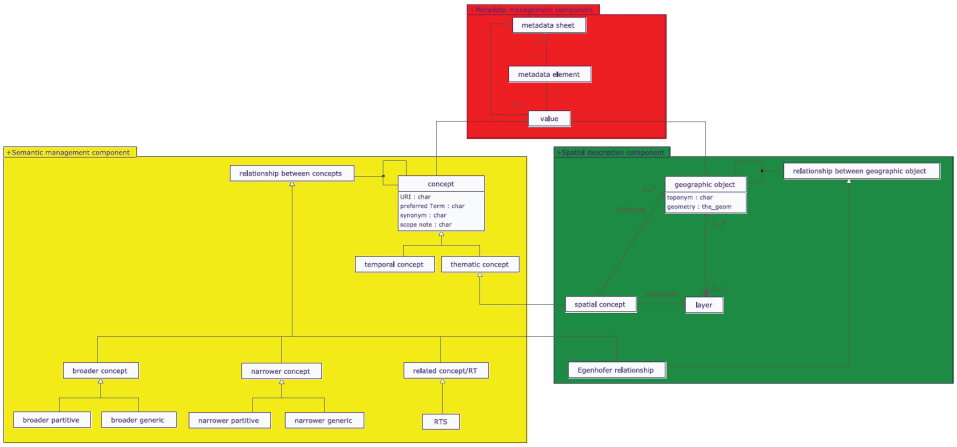

Fig. 5: Aggregation of generic components to control semantic and spatial descriptions (BARDE *et al.,* 2006).

gives additional details by making explicit the relationship between thematic and spatial concepts. We suggest the following link: "a spatial concept is a kind of thematic concept whose instances are geographic objects".

That way we can manage similar spatial and thematic descriptions (with terms or graphics) by reusing the same component for any metadata standard.

Ongoing implementation

A similar model has been used to generate physical data models (PDM) which enable the management of any metadata standard. Our current implementations are based on version 1.5 (or 1.6) of MDWeb open source software to set up an architecture made of different components:

- a multistandard and multilingual metadata cataloging tool implementing a generic approach[3] compliant with standardized implementations of metadata standards (XML Schemas, DTD...),
- a semantic and spatial descriptions manage-

3 like other softwares: NOKIS [8], M3Cat [BERKLEY *et al,* 2001], MetaCat.

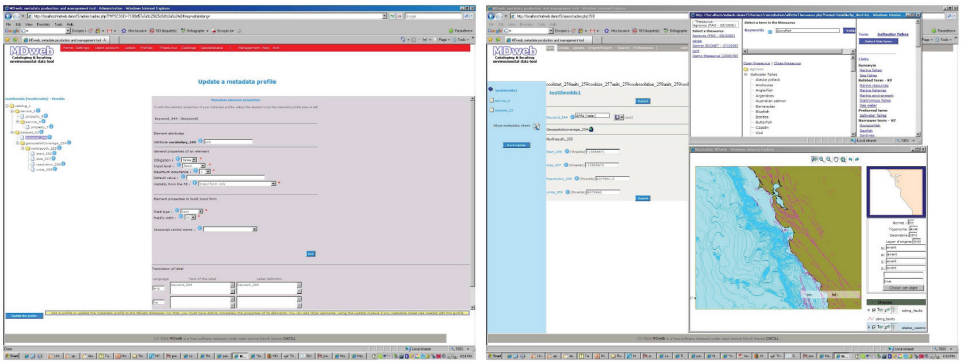

Fig. 6: Generic GUIs for metadata sheets edition.

ment tool compliant with (Web) Semantic standards SKOS (ISO 2788 / 5964) / RDF/ OWL and the main standardized formats for spatial information (GML, KML...).

- a three-tier (client-server) architecture with:
  - (1) a set of GUIs (inWeb browsers) integrating components to assist metadata edition or retrieval:
    - spatial description withWeb Mapping tools: Mapserver/ Mapbuilder,
    - thematic description with Controlled vocabularies tool: home made.
  - (2) Applications scripts: PHP[4] /Javascript/XML (with Apache Http server),
  - (3) Data storage: RDBMS: Postgres with Postgis (import of SKOS files into Postgres (by using JENA java API), XML repositories (metadata sheets...).

We used this architecture to import the XML schemas [W3C] of needed metadata standards into the physical data model (ISO 19139, SensorML, Thredds XML schema, SSDS, DCMI) by creating a new script. This script translates XML schemas statements into SQL statements[5] which fill the tables of our PDM. Once imported, users can handle any kind of metadata standard with a single GUIs set. This way, developments efforts focus on the same scripts whatever the standards.

4 a new Java release is currently used.

5 by handling XML nodes with the Document Object Model.

Homogeneous GUIs set

By using MDWeb 1.5 as a basis of development, we illustrated the first results and the interest of this methodology. This work has shown the relevance of such generic models to manage any metadata standards within a single architecture managing as well controlled vocabularies. The following GUIs have been generated by importing the XML schemas.

Import of any metadata standards and profiles edition

By importing a new metadata standard XML Schema through a generic script which translates this kind of formal specifications into its generic PDM, MDWeb can use the same scripts set to generate dynamically a GUI to set up profiles of this standard.

The GUIs shown in Figure 6 have been generated this way. The first GUI has been set up by interpreting a profile of Thredds metadata elements set (previously imported into the PDM). Similar GUIs have been generated with ISO 19139, SSDS or DCMI metadata elements sets.

Metadata sheet edition

Once a profile of an imported standard is set up, users can thereafter edit some instances (metadata sheets) of this profile. The second GUI of Figure 6 illustrates the form generated to edit an instance of a profile (Thredds metadata

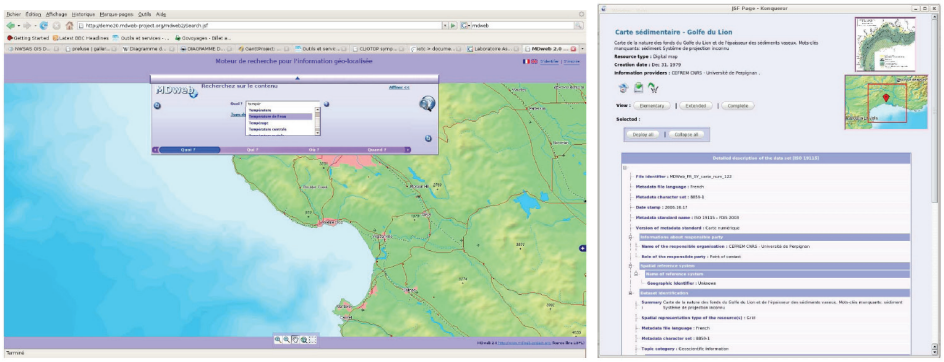

Fig. 7: Generic GUIs for metadata sheets searching and visualization.

standard). As thematic and spatial descriptions are usually part of any standard profiles (Fig. 3), we can assist (and control) any metadata sheet edition by reusing generic GUIs (such as pop-ups shown in Figure 6 for thematic and spatial descriptions).

Searching and visualization of metadata sheets

We present here GUIS related to another example of a generic conceptual model which is implemented in the next version of MDWeb (2.0, Java, not released yet). Being more compliant with an object approach, the underlying PDM improves the previous one and its ability to manage different metadata standards[6] . However an additional set of generic scripts (written in Java) is needed to handle this new PDM and to satisfy the same needs as the previous version (in terms of importing any standard, profiling it and implement it to edit metadata compliant with the related XML schema). So far, the ongoing implementation of this new version currently offers a GUIs set for the searching engine as well as for metadata sheet visualization (see snapshots in Figure 7, ISO 19115 in this case) and soon GUIs for metadata edition.

Conclusion

The case of informational resources (IR) management for Ocean Observatories illustrates general issues existing in many other domains: different information systems managing different metadata and data formats (with related services) to describe and treat the same kinds of IR. However, users still complain about difficulties to "get (or aggregate) the data" they need. New standards (summarizing the contents of previous ones and made explicit with XML schemas) for environmental and spatial data infrastructures interoperability (on Internet) are available. When implemented they facilitate IR sharing at a local or global scale. New kinds of information systems built on top of existing ones can cover a large part of user's needs. By using generic models and existing standards and open-source softwares, a single generic architecture, can manage:

- heterogeneous metadata standards (import, profiling, edition...),
- heterogeneous values: in particular controlled terms and spatial descriptions to describe core metadata elements,
- a common indexation table duplicating core metadata elements with homogeneous terms and values used by the search engine (CSW...),
- spatial IR described by metadata can be processed after being retrieved: either locally or

6 for now DCMI, ISO 19115/39 and SensorML schemas are managed but we didn't adapt the script to import a new XML schema into this PDM.

remotely by using Web Services with Web Browsers or/and rich clients (e.g. WMS / WFS /WCS...with Mapserver, Qgis, Udig...).

With such a system built on top of specific ones (managing specific standards) it is possible to comply with new reference standards without changing ongoing applications. Moreover, this approach facilitates the implementation of more sophisticated standards, like OGC CSW catalog interoperability standard, by concentrating the clients on a single server (based on a single scripts set) in charge of all the existing information systems behind. Data discovery can thus be improved, in particular queries expansion is made possible by using standardized semantic or spatial relationships (requests portability). Both user's and software engineer's tasks and needs are taken into account by a single scripts set.

Our implementations of this generic approach illustrates the interest of such an architecture but requires some reingeeniring of current models and scripts to deliver a fully operational tool. The MDWeb project aims to set up such a tool with the version 2. Developments efforts focus first on enabling the management of any metadata standard, profiling it, edit and/or search related metadata instances in different catalogs (with CSW). Syntactic interoperability is then made possible with other systems. However semantic interoperability has to be controlled by a dedicated component which is needed to make metadata content machine readable in order to match relevant metadata. This control is crucial to completely master the ability to expand queries towards different catalogs and understand the results sent back by external servers (whatever the terms or languages they use to describe their IR with semantic or spatial concepts).

References

BARDE, J., DESCONNETS, J.C., LIBOUREL, T. & MAUREL, P., 2006. Generic conceptual models for data and knowledge sharing. application to environmental domain. In *Hydroscience and Engineering, ICHE 2006, Metadata and Ontologies in HydroSciences, Philadelphie, USA, Drexel University*.

BARDE, J., DUANE, E. & DESCONNETS, J.C., 2007. A generic approach to manage metadata standards. *OSGeo*, 3:24-30.

BERKLEY, C., JONES, M., BOJILOVA, J. & HIGGINS, D., 2001. Metacat: A schema-independent xml database system. In SSDBM '01: Proceedings of the *Thirteenth International Conference on Scientific and Statistical Database Management, page 171, Washington, DC, USA*. IEEE Computer Society.

BERNSTEIN, P. A., HAAS, L. M., JARKE, M., RAHM, E. & WIEDERHOLD, G., 2000. Panel: Is generic metadata management feasible? In The *VLDB Journal*, p. 660-662.

CARON, J., 2004. Dataset inventory catalog specification version 1.0. http://www.unidata.ucar.edu/projects/THREDDS/tech/TDS.html.

GOMES, K.J., GRAYBEAL, J. & O'REILLY, T.C., 2006. Issues in data management in observing systems and lessons learned. In OCEANS 2006, p. 1-6.

ISO/TC 211, 2011. Geographic information. http://www.isotc211.org/.

KAZAKOS, W., VALIKOV, A., SCHMIDT, A. & LEHFELDT, R., 2002. Automation of metadata repository generation with xml schema. In in proceedings of *16 th International Symposium Environmental Informatics 2002 (Enviro Info), Vienna, Austria*.

OGC, 2011. Open geospatial consortium, http://www.opengeospatial.org/.

O'REILLY, T.C., HEADLEY, K., GRAYBEAL, J., GOMES, K.J., EDGINGTON, D., SALAMY, K.A., DAVIS, D. & CHASE, A., 2006. Mbari technology for self-configuring interoperable ocean observatories. In OCEANS 2006, p. 1–6.

WORLD WIDE WEB CONSORTIUM (W3C), 2004. Xml schema part 0: Primer second edition. http://www.w3.org/TR/xmlschema-0/.