Metadata catalogue :
thematic querying, data preservation and advertising

Cathy BOONNE[1], Anne Sophie ARCHAMBEAU[1], Vanessa MAIGNE[1], *e*-XMLMedia[2]

[1]IPSL, Institut Pierre Simon Laplace
Université Pierre et Marie Curie
T26, 4,place Jussieu
75252 Paris

[2]*e*-XMLMedia
31 Avenue du Général Leclerc
92340 Bourg la Reine
E_mail : Cathy.Boonne@ipsl.jussieu.fr

Abstract : A project of implementing a web application has been conceived at IPSL to generate automatically metadata documents and to treat user thematic requests. A meta database user interface has been developed and implemented using the web technology. Within the user interface, different tools are available to integrate metadata documents based on the DIF format. The e-XMLMedia Repository 2.0 manages the metadata documents in the meta database and allows users to access to the relevant information through thematic requests.

Résumé : Un projet proposant le développement d'une application gérant automatiquement la création de méta données et le traitement de requêtes thématiques a été spécifié à l'IPSL, pour répondre aux besoins des producteurs et utilisateurs de données. Une interface utilisateur accessible via Internet a été développée et implémentée, elle utilise la technologie web. Via cette interface, différents outils sont disponibles pour intégrer les méta données au format DIF. Les méta données référencées dans la base de données sont gérées par l'outil e-XMLMedia Repository 2.0, cet outil permet aux utilisateurs d'accéder aux résultats de leur requêtes thématiques.

1. Introduction

The "Centre de Données (former Pôle de Données)", belonging to the Institut Pierre Simon Laplace (IPSL/CNRS), coordinates and optimises the efforts to make the data and metadata available to the IPSL and external users. The IPSL Laboratories have stored all kinds of data, including model outputs, and campaign data with a large number of different types of observational data. The IPSL databases (BIPBIP, Climserv, Dataserv, Ether) are distributed in different servers; some of them are available from outside IPSL, directly and/or through the IPSL central server.

Metadata are playing a major role in data information preservation and advertising. In order to promote data through a metadata catalogue inside and outside IPSL, a project of implementing a web application interface has been conceived for:

- automating the generation of metadata in order to preserve information around the data production,
- treating thematic user requests through an automatic search engine.

2. The IPSL meta database concept

The IPSL meta database has been initiated in 1999 by the "Pôle Données" in collaboration with G. Gardarin, international expert about databases and XML (Extensible Markup Language) systems, to fulfil, at the same time, the needs of IPSL and of future European or international thematic networks.

2.1 Metadata format

Within the meta database context, the metadata catalogue has been generated by using the DIF (Directory Interchange Format) international standard format. The DIF has been chosen because it is a standard metadata format recommended by CEOS WGISS (Working Group on Information Systems and Services), whose expertise about information systems and services is of value for users of Earth observation data. By using the DIF, we can also take advantages of its compatibility with others standard metadata format like the FGDC (Federal Geographical Data Committee) standard ; then the conversion into the DIF is facilitated.

This format allows to create directory which describe well the data held in the different IPSL databases. It consists of a collection of fields which detail specific information about the data (for more information about the DIF, see http://gcmd.gsfsc.nasa.gov). This collection of informations are very useful and necessary for data users to identify and locate the desired data and in general for data preservation and advertising.
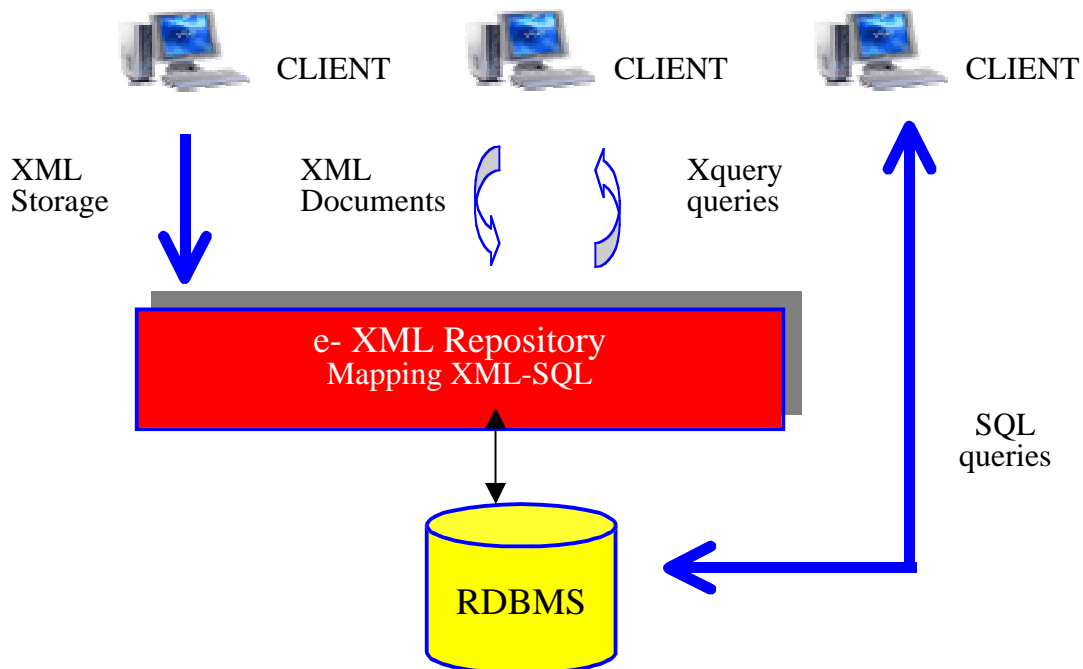
2.2 *e*-XML Repository 2.0

A new search engine : e-XML Repository 2.0 [1] based on XML, which is currently the leading language for representing and exchanging data , has been chosen to store and extract relevant information about data from the IPSL metadata catalogue. This engine has been developed by the e_XMLMedia society founded by G. Gardarin in collaboration with industrial partners.

The e-XML Repository 2.0 can be described as a middleware which is able to manipulate (insert, update) efficiently XML documents in a standard relational database. This repository provides also means to query the stored XML documents through a new generation of query language : Xquery [2]. These functionalities are made available through a set of JAVA/API (Application Programming Interface).

The documents may be provided with or without schema and DTD (Document Type Definition). The e-XML Repository 2.0 provides fast loading of document into the database and persistence of data. The document can be recomposed using queries and data can be searched through the use of XML structure, element of XML documents or data values. The main characteristics of this tool are :

- dual XML and XQL (XML Query Language) view of the database,
- all documents may be controlled by a schema and transformed into a relational schema,
- support of W3C standards like XQuery,
- maximum portability by the use of SQL, JDBC2, SAX2, DOM2, XQuery and a modular architecture (figure 1).

Fig.1 : Overview of the e-XML Repository 2.0 architecture

2.3 Meta database implementation at IPSL

The runtime requirements for e-XML Repository 2.0 are :

- JDK 1.3, the repository is built upon JAVA 2 platform,
- a servlet engine : Tomcat,
- Oracle 8i, database management system accessed through a JDBC 2 driver.

At IPSL, we have installed these different tools on a single machine (PC/Linux, Redhat 6.2). Within e-XML Repository 2.0, a web-based administration interface allows the creation of the different repositories, the loading of XML documents with or without schema , as well as the updating and deletion of them. Using it, we have partially integrated a first catalogue of metadata.

Besides the web-based administration interface, a user interface has been developed using the web technology and implemented [3]; we have plugged the e-XML Repository 2.0. on it.

3. The IPSL meta database user interface
3.1. Facilities

We have developed the user interface for two types of users :

- data and metadata scientific producers,
- data and metadata scientific users.

To serve the first one, IPSL has developed a web-package for editing, updating and deleting the DIF metadata documents. After the DIF validation, the captured DIF information are converted into XML documents and automatically integrated into the meta database. In a near future, the DIF capture will be done using a XML template.

For the second type, data users, a dedicated website allows them to query thematic requests in a user-friendly way. Following the metadata collections tree, they can explore easily the meta database and its content.

3.2 Overview of the user interface

The user interface is composed of two main parts : one for data and metadata producers (A) and the second part for data users (B). The figure 2 shows a view of this interface.
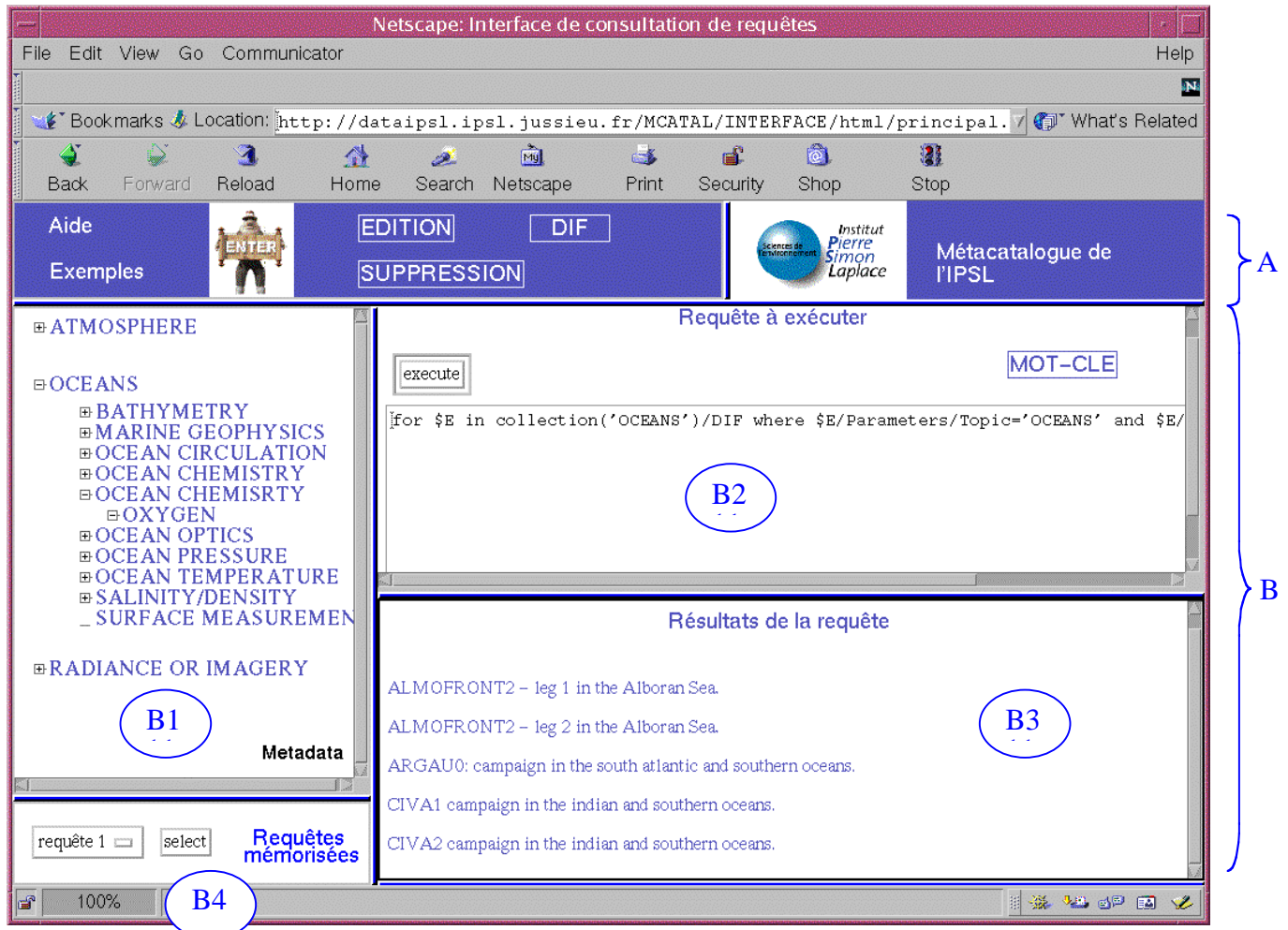

Fig.2 : View of the user interface

The part on the top (A) allows a metadata producer to fill a DIF template, update or delete it (see button DIF). This part is not public, producers can access it using an identifier. On the contrary, the part on the bottom (B) is free. This part B is composed of four frames. The big part on the left of the screen (B1) displays all existing stored collections (e.g. on the screen : ATMOSPHERE, OCEANS, RADIANCE OR IMAGERY) in the meta database. It is possible to expand all of those collections to explore them. Once the relevant data are identified, the associated request is directly written on the frame B2 in a transparent way for the user. The user has to click on the button "Execute" frame B2 to send his thematic query ; the titles of each DIF documents, result of the corresponding request, are listed on the frame B3. By clicking on a result title, a window is open with the DIF content. Thematic requests can also be memorized using the tool available in frame B4.

3.3 Towards the data retrieving

In order to achieve this IPSL internal project, it will be very interesting through this user interface to retrieve not only metadata information but also the relevant data by developing wrappers adapted to each kind of data structure. e-XMLMedia has developed a product : Mediator which provides the required functionalities to query, in a uniform way, heterogeneous data sources distributed on various sites

interconnected through the Internet. The next step of the IPSL project is to implement and test this product plugged on the user interface we have developed.

4. Conclusion

Currently, the IPSL meta database user interface is completely developed and implemented but a test period is necessary to check all the functionalities integrated. The operational site will be accessible in autumn 2002 (http://dataipsl.ipsl.jussieu.fr/metadata).

References

[1] e-XMLmedia. *e-xml Software Suite, Repository 2.0 tutorial*, vol.3, 2002.
[2] e-XMLmedia. *e-xml Software Suite Query Language*, 2002.
[3] V. Maigné. *Contribution à l'implémentation d'une interface de requête sur une base de données thématique*, 2002.