XML-based Metadata Management for Biological Data

J. L. Campos dos Santos$^{1,2}$ and R. A. de By$^2$

Abstract

Amazon biological data have been collected, primarily by single or small group of researchers in small areas, over relatively short periods of time. Questions on “ecological patterns and process variation in time and space, and what are the causes and consequences of this variability” are still important. To answer these questions, far more data are required than what could be collected, managed, and analysed in a single organisation. Researchers need detailed documentation to accurately interpret and analyse historic or long-term data sets, as well as, data from complex experiments. This requires understanding of the type of variables, the units adopted, potential biases in the measurement, sampling methodology and a series of facts that are not represented in the raw data, but rather in the metadata, which is non-existent for the majority of data sets.

We have implemented a solution to manage biological metadata; supported by the FGDC metadata standard. It includes the use of XML-based metadata representation that has been incorporated into client/server architecture. This allows users to access data and metadata over the web. The digital management of metadata will give the users the possibility to locate and understand data through time.

1. Introduction

Biologists at universities, museums, and governmental institutions collect biodiversity data from isolated surveys. They cover small geographic areas over short period of time (Stockwell, 2000). In the Amazon a considerable amount of environmental information has been originated from biological collection data. For this particular region, data are abundant and most of them are kept in biological collections, which are constantly growing. This dynamic process presents some major problems that affect data management, analyses and consequently the information dissemination,

$^1$ The National Institute for Amazon Research, André Araújo Avenue 2936 – Petrópolis, 69.083-000 - Manaus – Amazonas – Brazil. Tel. +55 (92) 643-3032, Fax: +55 (92) 643 3311; Email: lcampos@inpa.gov.br, http://www.inpa.gov.br

$^2$ International Institute for Geo-Information Science and Earth Observation, Hengelosestraat 99, P.O. Box 6 - 7500 AA Enschede – The Netherlands. Tel. +31 (53) 487 4374, Fax: +31 (53) 487 4335; Emails: {santos, deby}@itc.nl, http://www.itc.nl
that is, data are scattered around institutions, no metadata standard has been adopted and data are exchanged amongst researchers in close range with data producers with no easy access to the data itself.

Despite of the success of network communication in connecting scientific institutions, only small parts of its environmental information are available on the World Wide Web. The use of web to provide scientific information is progressing in phases, from a mere presence, to information and database access, to nodes on exchange, and to information portals3 (Niemann 2000). An information portal can provide access to all information on the Web. However, scientific audience requires a more specific set of resources that can lead to a coherent view of information gathered from disparate sources (White 1999). By adopting the Enterprise Information Portals approach; which focus on a specific domain, and enable facilities to deal with collaborative and decision processes in organizations, the problem of disparate scientific information can be alleviated (Goldfarb and Prescod 2002). Metadata descriptions depend on a formal description of data, which can be of semi-structured form. The best option is the use of the Extended Markup Language (XML), since XML was designed to deliver structured content over the web (Harold and Means, 2001). The strength of XML technology is that it can deal with semi-structured data, which are not suitable to be managed by (object) relational database technology (Campos dos Santos et. al., 2002). The development of XML-based metadata solutions that makes biological metadata accessible is urgent needed.

Some work has been carried out by groups, such as the National Partnership for Advanced Computational Infrastructure (NPACI) and San Diego Super Computer Center (SDSC), which is centred in the field of environmental informatics and aim the development of common information infrastructure for research and education (NPACI 2002 and SDSC 2002). Tools such as The Species Analyst (TSA 2002) and the XML-based tools for Data-Intensive Computing Environments – (DICE 2002), integrates data from many database, independent of platform, software technology and data formats and displayed, on the user Internet browser.

This paper presents a solution for metadata management, which was implemented for the “Instituto Nacional de Pesquisas da Amazônia” (INPA), Manaus-Amazonas, Brazil. The solution applies XML technology to manage and distribute metadata. It is implemented in the Web environment, where users can access data and metadata by querying a XML repository and searching data for download. We focus on scientists as our prime users who are able to describe metadata. The FGDC metadata

---

3 A portal, web site or service offers a broad array of resources and services, such as e-mail, forums, search engines, and on-line business. The first Web portals were online services, such as America Online, that provided access to the Web, but now most of the traditional search engines have transformed themselves into Web portals to attract and keep a larger audience.
standard, which includes de Biological Data Profile, was adopted to be the INPA’s metadata solution (FGDC 2002).

We described the FGDC standard as an XML schema and automatically generates a XML template. We use the BioME portal to host components (XML file, schema, XML Editor and full documentation) to be deployed by request from users at Online Local Nodes (OLNs). Users of an OLN can use the XML Editor to upload the template and describe scientific data sets. Each metadata description corresponds to a data set that can be submitted for storage. The metadata is kept in an XML repository (the XYZFind Server System) and can be widely accessed via XYZ Query language4 (XYZFind 2001). The repository accepts any number of XML metadata descriptions and maintains a single data representation of all the XML metadata it receives. The metadata can be retrieved by users and can be updated, or removed from the repository by the BioME portal manager. Once the repository is indexed, search and query operations are available for global access.

2. **Metadata Management**

Most scientists have a feeling for data files, however, for large number of professionals in the field of environmental sciences, the concept of metadata is not so clear. Metadata comprehends information about where, who, when, why, etc. It is the short, searchable form of documentation used to find and understand data. Metadata can also be regarded as card catalog information. The non-existence of metadata descriptions can make data sets unusable and impossible to be shared, and also its information disseminated.

Metadata aim to organize and maintain data by alleviating the phenomena of information entropy5. It can contribute with information to data catalogs and clearing-houses, and provide essential information in case of data sharing. It describe several aspects of data, including: data set identification, description of its quality, the spatial organization and reference, the properties, the way to distribute the data and the metadata reference (FGDC 2001).

After the FGDC Standard had been approved in 1994, it underwent significant enhancements by allowing the metadata producer to profile the base standard by defining a subset of the metadata entities and/or elements that are used by a specific discipline or organization. The development of the FGDC had provided essential elements to research groups at INPA that after evaluation of the standard had decided to adopt it for describing biological data and to implement it across the INPA’s biological collections.

---

4 XYZFind Server System and XYZQL Language is trademark of XYZFind Corp – Kirkland Way, Suite 310, Kirkland, WA 98033, USA.

5 Phenomena that occur during the normal degradation of information content associated with data and metadata over time (Michener et al, 1997).
The evaluation focused on the properties defined for the Biological Data Profile that can well describe data sets with species classification and determination information. Regarding biological data dissemination and metadata preparation, in recent scientific workshop held in Belém - Pará, Brazil, it was discussed the effectiveness and influence of this standard in other research institutes in the Amazon, specially those which hold information of similar nature. The result indicates that there is willingness to adopt the standard amongst institutions.

FGDC Metadata Schema

All the entities and attributes described in the FGDC metadata standard were mapped into the W3C XML schema definition\(^6\) using the XML Spy Suite system\(^7\). The schema is an XML language for describing and constraining the content of XML documents.

Based on the schema, the system generates an FGDC-based XML output (a well-formed XML document). Users by using an XML Editor can upload the template for insert information. This guarantees the integrity of the final version of the XML file that is later submitted for registration and to be stored in a XML repository. The Figure 1 presents the way the FGDC XML Template was produced. Figures 1 and 2 present the way XML template was produced and an example of the XML template respectively.

3. Implementation: Three Tier Architecture

We have implemented a Client/Server architecture that is comprised of three tiers: Client Environment, Web Server and Database Servers. This approach presents advantages regarding modularity and easy the implementation of load balancing. Figure 3 presents the infrastructure implemented that allows data and metadata management.

The infrastructure is comprised of the following components:

- **Client Environment** - The infrastructure supports two types of users: Providers and Brokers\(^8\). The providers are those users or agents who are responsible for data collection. The brokers are those who use the raw data to extract knowledge from it and also describe metadata. Both users operate at Internet and Intranet levels. At the Internet, the users can access and perform operations available at the BioME portal supported by the Apache Web Server, the XML Repository via the XYZFind Server operations and at any active OLN. The main operations at this level include: downloads (data sets specifics, documents

\(^6\) W3C XML Schema is a W3C Recommendation; In http://www.w3.org

\(^7\) Spy View Suit 4.3 is a trademark of of Altova GmbH, Rudolfsplatz 13a/9, A-1010, Austria.

\(^8\) We use similar definition of users as Tomasick and Simon (Tomasick and Simon 1997).
and application), queries - that can be submitted to a database or the metadata management system, browser and servers specifics operations. At the Intranet level (Active OLN), users have similar functionalities as a client environment.

- **INPA Internet/Intranet Backbone** – The INPANetwork is supported by the ATM (Asynchronous Transfer Mode) technology; and is comprised of sub-local area networks with fast backbone. At the moment, the uplink to the Internet pre-

---

9 Users can describe metadata via Web. However, the process is time consuming and due to local networks and Internet bandwidth performance, users can describe metadata in local machines.
sents limitations that can affect the overall performance of this infrastructure due to the data/metadata volume growth.

- **Web Server** – Apache HTTP Server was installed to allow access to the BioME Portal. Apache has been the most used web server on the Internet, reaching 56% of dominance (Apache 2002; Netcraft 2002).

- **BioME Portal: Deployment of Components** – A portal has been implemented to provide access to the BioME components, that is, full documentation of FGDC standard, the XML template and a XML Editor, as well as to permit online editing of metadata. Editing a metadata is time consuming and due to precarious bandwidth speed the performance of this process would be affected. The best option would be to allow the deployment of the XML template and an XML Editor. The Editor, once installed in the client environment, uploads the template to allow editing. This guarantees that the final product is a well-formed XML document; fit to be uploaded into the XML Server.

- **XML (Database) Server: Metadata Repository** – We have selected the XYZFind Server, which is an integrated repository, search, and query server designed for XML. The system is easy to deploy and with sufficient functionalities for XML management. Combining the best features of databases and full-text engines, XYZFind delivers robust storage and retrieval access without dependence on schema design, transformation, or DTDs. Unlike traditional database technologies that decompose and store XML in document specific tables, XYZFind eliminates schema design and data mapping. It achieves this by decomposing XML documents into a single, schema data representation. This approach provides persistence, and is comprehensively indexed, that is at the level of word and numeric value, and all associated path information). XYZFind makes storing XML documents simple and delivers the high performance structured query you expect from a database, and enables seamlessly integrated keyword search.

4. **Conclusions**

We have presented a solution for manage XML-based metadata for biological collection data. The FGDC, a consolidated metadata standard, has been adopted and implemented as template to be deployed by request for editing in stand-alone mode. The solution has been incorporated into a client/server infrastructure that allows metadata and data dissemination for users in a global scale via the Web.

Taking into account the Scientific Collection Program (an INPA’s initiative), this work can contribute for data/metadata exchange, particularly across the Amazon, bringing to light information that are essential for preservation and for the sustainable development of the region as well as documenting the existent data sets, a clear sign that information chaos can be controlled.
This work is particularly interesting for developing countries since the system developed and tools used are either public domain or from Open Source, adding low cost for a robust solution to any particular institutional. To complement this work, a module for producing dynamic report outputs at the web interface level needs to be implemented. Users are now able to query/search for metadata descriptions and when found, they must be able to produce user-tailored outputs, such as HTML, PDF etc.

**Acknowledgement**

We would like to thank our colleagues Wim Bakker and Arta Dilo for their comments. Thanks are also due to Sidineia Amadio for providing metadata descriptions of a fish data set from Uatumã River, Amazon-Brazil.

**Bibliography**


