FACILITATING INTERDISCIPLINARY SCIENCES BY THE INTEGRATION OF A CLOSi-BASED DATABASE WITH BIO-METADATA

J. L. Campos dos Santos [1,3], R. A. de By [1] , P. M. G. Apers [2] and C. Magalhães [3]

1. International Institute for Geo-Information Science and Earth Observation (ITC), Hengelosestraat 99, P.O. Box 6, 7500 AA, Enschede, The Netherlands – (santos, deby)@itc.nl
2. University of Twente, P. O. Box 217, 7500 AE, Enschede, The Netherlands - apears@cs.utwent.nl
3. The National Institute for Amazon Research (INPA), André Araújo Avenue 2936 – Petropólis, 69.083-000 - Manaus – Amazonas – Brazil – (lcampos, celiomag)@inpa.gov.br

Commission IV, WG IV/2

KEY WORDS: Database schema, Metadata, Biodiversity Information Systems, Web Environment

ABSTRACT:

Biodiversity information has been collected and compiled during many unrelated and independent projects across Amazon region. Institutions on their own maybe unable to answer crucial questions, as their answers may depend on a multi-disciplinary context. Since their situation is still considerably isolated, some solutions adopted impose redundancy leading to high costs. The use of computer technology has been a fundamental resource for biological information management. However, such information is heterogeneous and to understand and automatically manage it accurate schema representation and formal metadata description are needed. This paper presents concepts of the CLOSi schema, which was conceived to ease and stimulate the development of biological collection databases. The schema represents functional groups that are specified in terms of object classes and their relationships. Another important issue presented, is the management of good quality bio-metadata that can be integrated with the data. For that, we implemented a solution to gather, manage and disseminate biological metadata, which is integrated with our CLOSi-based database. The FGDC metadata standard was adopted and represented as an XML schema. The schema was mapped to a well-formed XML template. A client/server infrastructure was tested to manage and disseminate data and metadata about Amazonian biodiversity.

1. INTRODUCTION

A vast amount of information has been compiled on the properties and functions of the Earth's living organisms, and an increasing proportion of this information is contained in large repositories in digital form. These include biodiversity data on the distribution of plants, animals and microbes; detailed genomic maps; compilations of the physiological functions of organisms, and information about the behaviour and function of species within ecosystems. Because these data have been collected and compiled during many unrelated, independent projects, their full research potential has not been reaped.

In the Amazon, even though the region is considered to have one of the highest concentrations of biodiversity, the data situation is critical. One of the most important assets are the biological collections, which are distributed across institutions. Biological collections are represented by samples, as well as by associated scientific information. The challenges that institutions are facing is to organize biological data in a Biodiversity Information System (BIS) with a consistent and efficient database implementation that allows the use of legacy data, and supports integration, intelligent and easy access for dissemination and sharing purposes.

To identify users and system requirements in this context, institutions were visited and problems identified in four domains: information production method, source of biodiversity data and information, degradation of information and inherent biodata modelling.

The "Instituto Nacional de Pesquisas da Amazônia" (INPA), is promoting computer-based solutions to help interdisciplinary collaborations and scientific advances. This aims to contribute to biodiversity preservation and sustainable development. The current interest of INPA is to share and disseminate data and information from its assets for a global scale audience. It started by implementing a strategy for information delivery, which includes: coordination, integration, analysis, presentation and delivery. To achieve that, understanding of the physical construction of its biological collections and scientific datasets is required. It also requires the knowledge about the data and objects information that will eventually be stored in the database. To identify users and system requirements, interviews and an evaluation were conducted.

The results of the study were grouped and clustered as schema objects, representing the items in collections. The schema provides a mechanism for database definitions that can be implemented in a (object) relational database management system.

Complementary to this, there was the need to describe and manage metadata and to integrate them with the data. We adopted the extended FGDC metadata standard, which incorporates the Biological Data Profile. Metadata has two purposes: (1) To help search and retrieval, or identify the location of data that meet a user's selection criteria, e.g., the "card catalog" type of descriptive information about a data set, and (2) To help a user to fully understand the data content and evaluate the usefulness of the data for his or her purposes, i.e., data set documentation.

This paper presents an implemented computer-based infrastructure for biodiversity data and metadata management. We focus on the support of CLOSi schemas for database implementations and the metadata standard for biological profiles. The standard has been implemented as XML schemas and transformed to an XML template. CLOSi databases can be integrated to Bio-Metadata through standard attributes and user-defined attributes. This method of integration allows users to access metadata information when querying collection objects, and to access data sets when browsing metadata descriptions.

2. SCHEMA REPRESENTATION FOR BIOLOGICAL COLLECTIONS

The design of database systems for managing biological collections requires the understanding of the physical structure of these biological collections. It also requires knowledge about the data and its characteristics. To acquire that, users were involved in the processes of data and systems requirements, since they play an important role, especially during data requirement analysis (Sonderegger *et al*. 1998; Campos dos Santos *et al*. 2000). At INPA, this phase was carried out in two ways: Interviews and descriptions evaluations.

Each participant in these processes was a specialist in some taxonomic group or a certain biological aspect of some taxonomic group. The interviews had an open format and researchers were asked the same general questions. This allowed us to elaborate a protocol to conduct the steps from field sampling to record information.

The data collected during a field mission consist of two parts: a general one, which holds the information that is normally collected in all studies (e.g., date, time, locality description), and a specific one, that corresponds to the scientific interest of a study (e.g., the altitude of a locality or the moon-phase may be of interest in one study but not in another). Interviewing scientists that worked on different studies and in different institutions helped to differentiate between information that is common to all and that which is used by just a few scientists. The result of the interviews was grouped by functionality and clustered as object types, representing the items in collections. The clusters were organized as: Collection Management, Collecting Event of Collection, Locality of Biodiversity Data, Taxonomy, Agent of Collection and Reference. This schema is called CLOSi, and clusters and their relationships are presented in Figure 1.

In the following, we present the main concepts of CLOSi; which include: clusters, object classes, relationships, specialization, attributes and its constraints, control value classes and notation.

CLUSTERS

A cluster groups a set of inter-related object classes. This forms the structure of a schema that can be used for the development of a database system. The schema combines a number of clusters and represents functional groups that are specified in terms of object classes and their relationships.

OBJECT CLASSES AND RELATIONSHIPS

An object class describes a population of objects in a collection and is identified by a unique name. One object class can have a relationship to other classes of any cluster. An object class is described by a class description and is associated with one or more attributes. A class can be a specialized class and for that, the class relationship is of type Is_a. An object class must belong to a unique cluster and can be associated with an optional list of control value classes (set of ordered atomic values).

IS_A OBJECT CLASSES

Specialization is an abstraction method that allows definition of object classes describing a subset of the population of another object class (super class). The inverse of specialization is generalization.

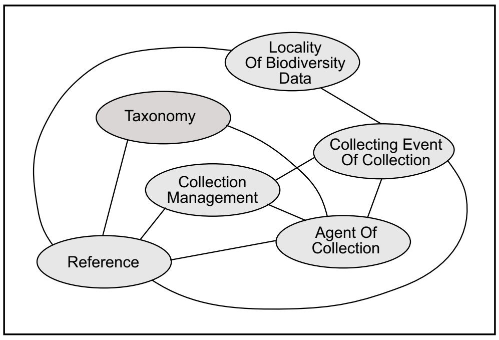

Figure 1 – CLOSi Clusters and their Relationships

ATTRIBUTES

An attribute is the basic unit of information about an object class occurrence. It may be local to the object class or inherited by relationship. CLOSi supports four types of attributes:

- **Standard** -- This attribute describes the common property of an object class, that is, attribute name, cardinality, the attribute data type and the attribute description. An object class can also be an attribute data type.

- **Composite** -- A composite attribute is formed by the concatenation of one or more attribute members and an attribute description.

- **Derivation** -- A derivation attribute represents a relationship between object classes and indicates that the attribute has an original object class (derivation_class_from), which derives to be an attribute of another object class (derivation_class_to).

- **Geo_attribute** -- It describes a geo-referential characteristic of the object-classes. It is described by an attribute name, cardinality and a Coordinates type or Distance. Coordinates of latitude and longitude are each defined by four values: degree, minutes, seconds and hemisphere. The Coordinate type can be of geographic, rectangular or nodes type. The geographic types describe the latitude and longitude and are also described by four values: degree, minutes, seconds and the identification of the hemisphere (e.g., N, S, E or W). The rectangular type also describes the latitude and longitude, which are defined by two values: rectangular co-ordinates in meters and the hemisphere. The nodes type defines the nodes of a chain.

ATTRIBUTE CONSTRAINTS

An attribute associated with an object class has a cardinality constraint, which specifies the minimum and maximum number of values for attributes. In the absence of a specified cardinality constraint, it is by default `[0,1]` for single value attributes, and `[0,]` for set-of value or list-of valued attributes, where an unspecified maximum represents and unlimited number of values. Also, the CLOSi schema assumes referential integrity regarding attributes associated with controlled value classes: the value of such attributes must belong to their associated value classes.

CONTROLLED VALUE CLASSES

Each controlled value class defined in CLOSi has a unique name, and can either be of string or numeric type. A controlled value class can be associated with an optional standard value, a class description, the name of cluster and the class where the controlled value is described and a list of declared-domain properties. Declared-domain properties are defined as a tag-value pair and will be treated by application programs.

A string type controlled value class consists of a set of enumerated atomic values, which are all strings. If a standard value is declared, it must be one of the enumerated atomic values. Each string type of controlled value is associated with a `CODE_TYPE` associated with a data type for this code. Data type can be numeric type like: `INTEGER, SMALLINT, TINYINT, REAL, FLOAT, DECIMAL` and `NUMERIC`, or character string type like: `CHAR` and `VARCHAR`. A code can consist of alphanumeric characters and special symbols (. : + - * / \ ! =). The symbol "-" is allowed only when the code starts with A-Z or a-z.

Each controlled value also has a value description indicating for which cluster and class the controlled value class has been defined, as well as a standard value and a description of the controlled value class.

A numeric type control value class consists of a set of ranges, where each range is either a number or an interval defined by a lower and upper limit. For example[1], a numeric type controlled value class `INTERVALS` can be defined as a `1-10, 100, 200-300`. If a standard value is specified it must be within ranges defined for this controlled value class. Numeric type controlled values or range can have value descriptions but cannot be associated with a `CODE_TYPE`.

CLOSI NOTATION

A cluster is graphically represented by a double solid line rectangle with its name placed in the middle of the internal rectangle. There are two ways to represent classes: classes that belong to a current cluster (part of a current cluster) and those belonging to an external cluster. The first is graphically represented by double rectangle with dashed line in the external part. The latter classes are represented by a single rectangle with solid line.

Relations can exist between a cluster and its classes (cluster association), between classes to specify a derived relation, represented as solid arrow and between classes to define Is_a relation, represented as a dashed arrow.

Classes have attributes that are graphically represented as solid arrow. It is represented by attribute name and data type, including control value class and coordinates type.

The composite attribute is represented as a rectangle marked by two bullet point attachments to the attribute members. The attribute members are described as normal attributes.

2.1 CLUSTER TAXONOMY: AN ILLUSTRATION

To partially illustrate the CLOSi schema, we present in Figure 2, the cluster Taxonomy and its relationships. The cluster describes information about the taxonomic classification, identification and the ecological relations of the taxons. The object classes `Taxon_Name`, `Taxon_Relation`, `Classification` and `Determination` are the classes that describe the cluster. The object classes `Collection_Object`, `Agent` and `Reference_Work`, belong to external clusters and have relationships to classes within cluster Taxonomy.

The object class `Taxon_Name` describes the taxon at the lowest possible taxonomic level. The attributes of this class include: `Relation` (a list of ecological associations to other taxa), `TaxonAuthors` (the agents, persons, that made the original description of the taxon), `Repository` (the agent responsible for the specimens), `Synonym` (a list of names used for this taxon.), `Parent_Taxon` (the taxon rank that lies in the Linnean hierarchy immediately over this taxon), `TaxonOrigRef` (the reference that contains the original description of this taxon), `References` (the references in which the taxon is mentioned), `Classific` (the association of the taxon name with classification attributes). Additional attributes that describe this class include: `CommonNameandRegion` (a list of the common name of the taxon with its respective region), `Rank` (the Linnean rank of the taxon, e.g., Family, Order, Subfamily, etc), and `GeneralTaxonCoverage` (the geographic range of a taxon).

There are specific attribute relationships from classes within the same cluster to object class `Taxon_Name`. From `Determination`, there is the attribute `Taxon` (the taxon that represents the result of the determination); and from `Taxon_Relation`, there is the attribute `Relation_Host` (the taxon to which this relation exists).

Also, there are attribute relationships amongst object classes from the Taxonomy cluster to external object classes. The object class `Classification` defines a classification of the taxon in a defined structure and has four attribute relationships: `Author` (the agent - person, that made the original description of the classification); `References` (the references in which the classification is mentioned); `ClassificationOrigRef` (the reference that contains the original description of this

[1] This example was partially described in (Chen & Markowitz, 1996).

classification); and UpdatedBy (a default object class attribute).

The object class Determination describes the information about the process of the determination of the taxon name. It has the relationship attribute CollObject (the identifier of exactly one collection object for which the determination has been made); Determiner (the list of agents that made de determination); and the UpdatedBy (a default object class attribute).

All CLOSi object classes have three default attributes: Created (date and time of the creation of the object), Updated (when the object was last updated) and UpdatedBy (identify the agent, in this case, the person that made the last update of the object).

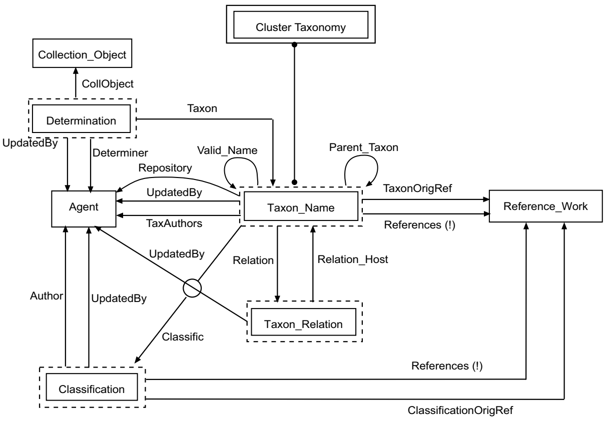

Figure 2 – Cluster Taxonomy and its Relationships

EXAMPLE OF CLOSI DESCRIPTIONS

This example uses real data from INPA's Ichthyology Collection. The information presented corresponds to the object classes and relationships of Figure 2. The Figure does not present the list of attributes of object classes.

Object Class Taxon_Name
Attribute CommonNameandRegion: *Amazonicopeus elongates*
Attribute Relation: Parasitism
Attribute Repository: INPA - Ichthyologic Collection
Attribute TaxAuthors: V. E. Thatcher
Attribute ParentTaxon: *Amazonicopeidae*
Attribute TaxonOrigRef: V. E. Thatcher, 1986
Attribute References (!): V.E. Thatcher,1991; M.A.P.M. Amado, 1995
Attribute Classific: Arthropoda, Crustacea, Maxilopoda, Copepoda, Poecilostomatoida, Amazonicopeidae, Amazonicopeus, Amazonicopeus elongatus

Object Class Taxon_Relation
Attribute RelType: Parasitism
Attribute Remarks: This taxon is a parasite of PESCADA fish (*Plagioscion quamosissimus)* and the place of fixation at the gill arches.
Attribute RelationHost: *Plagioscion squamosissimus*

Object Class Classification
Attribute Name: Arthropoda
TypeTaxonRankName: Phylum
Atribute Name: Crustacea
TypeTaxonRankName: Subphylum
Attribute Name: Maxilopoda
TypeTaxonRankName: Class
Attribute Name: Copepoda
TypeTaxonRankName: Subclass
Attribute Name: Poecilostomatoida
TypeTaxonRankName: Order
Attribute Name: Amazonicopeidae
TypeTaxonRankName: Family
Attribute Name: Amazonicopeus
TypeTaxonRankName: Genus
Attribute Name: Amazonicopeus elongatus
TypeTaxonRankName: Species
Attribute ClassificationOrigRef: K. Bowman
Attribute Author: K. Bowman
Attribute UpdatedBy: C. Magalhães

Object Class Determination
Attribute CollObject: *Amazonicopeus elongatus*
Attribute Taxon: *Amazonicopeus elongatus*
Attribute Determiner: V.E. Thatcher
Attribute Date: 15-07-1986
Attribute UpdatedBy: C. Magalhães

Object Class Agent [In Cluster Agent of Collection]
Attribute FistName: V.
Attribute MidName: E.
Attribute FamName: Thatcher

Attribute FistName: C.
Attribute MidName:
Attribute FamName: Magalhães

Object Class Refence_Work [In Cluster Reference]
Attribute Author: T. Bowman; L. G. Abele
Attribute Title: Classification of the recent Crustacea. P. 1-27. In Abele, L.G. (ed.), Systematics, the Fossil Record, and Biogeography. New Yourk, Academic Press. 319 pp. (The Biology of Crustacea, v.1).
Attribute Year: 1982

Attribute Author: V.E. Thatcher
Attribute Title: The parasitic crustaceans of fishes from the Brazilian Amazon, 16, *Amazonicopeus elongates* gen. Et sp. Nov. (Copepoda: Poecilostomatoida) with the proposal of Amazonicopeidae fam. Nov. and remarks on its patogenicity. Amazoniana, 10(10):49:5.
Attribute Year: 1986

Attribute Author: M.A.P.M. Amado; J-S. Ho; C.E.F. Rocha
Attribute Title: Phylogeny and biogeography of the Ergasilidae (Copepoda, Poecilostomatoida), with reconsideration of the taxonomic status of the Vaigamidae. Contributions to Zoology, 65(4): 233-243.
Attribute Year: 1995

3. METADATA STANDARD FOR BIO DATA

The Federal Geographic Data Committee (FGDC) coordinates the development of the National Spatial Data Infrastructure (NSDI). The NSDI encompasses policies, standards, and procedures for users to cooperatively produce and share geodata. Metadata or "data about data" describe the content, quality, condition, and other characteristics of data. The FGDC had approved the Content Standard for Digital Geospatial Metadata. The objectives of the standard are to provide a common set of terminology and definitions for the documentation of digital geospatial data. The standard establishes the names of data and associated elements to be used for these purposes, the definitions of these compound elements and data elements, and information about the values that are to be provided for the data elements (FGDC, 1998).

In 1999, the FGDC Steering Committee endorsed the Biological Data Profile of the Content Standard for Digital Geospatial Metadata (CSDGM). The purpose of this standard is to provide a user-defined or theme-specific profile of the CSDGM to increase its utility for documenting biological resources data and information. This standard supports increased access to and use of biological data among users on a national (and international) basis. It also helps to broaden the understanding and implementation of the FGDC metadata content standard within the biological resources community. This standard also serves as the metadata content standard for the National Biological Information Infrastructure (NBII). The extended standard can be used to specify metadata content for the full range of biological resource data and information. This includes biological data, which are explicitly geospatial in nature, as well as data, which are not explicitly geospatial (such as data resulting from laboratory-based research). This also includes "information" categories, such as research reports, field notes or specimen collections (FGDC, 2001).

3.1 INTEROPERABILITY OF BIO METADATA

XML, the Extensible Markup Language, provides an intuitive method for data structuring and organisation, based on tags written in text files. XML allows to set standards, defining the information that should appear in a document, and in what sequence. XML, in combination with other standards, makes it possible to define the content of a document separately from its formatting, making it easy to reuse that content in other applications or for other presentation environments. Most important, XML provides a basic syntax that can be used to share information between different kinds of computers, different applications, and different organizations without needing to pass through many layers of conversion.

XML provides a simple format that is flexible enough to accommodate diverse needs. Even developers performing tasks on different types of applications with different interfaces and different data structures can share XML formats and tools for parsing those formats into data structures those applications can use. It offers to users many advantages, including: simplicity, extensibility, interoperability, openness, and a core of experienced professionals. XML operates on two main levels: first, it provides syntax for document markup; and second, it provides syntax for declaring the structures of documents.

XML is derived from the Standard Generalized Markup Language (SGML). SGML has found its main customer base in organizations handling large quantities of documents. SGML's development provides the foundations for XML, but XML has a smaller and simpler syntax, targeted at Web developers and others who need a simple solution to document creation, management, and display.

The features of XML indicate it is suitable language for metadata representation. We use the XML schema for the definition of a BioMetadata XML file template implementing the biological standard profile, allows representing all the characteristics of the metadata elements like hierarchy, repeatability, optional or mandatory items, selection, attributes, data types and values based on integrity constraints. This processs allow schema validation to ensure well-formed XML file outputs. The Figure 3 presents the node BioMetadata schema, implemented from the FGDC standard specifications and mapped to a XML template.

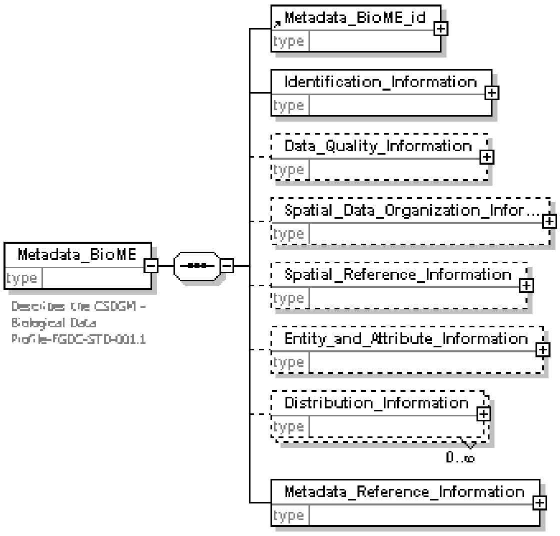
Metadata_BioME
type
Describes the CSDGM - Biological Data Profile-FGDC-STD-001.1
Metadata_BioME_id
type
Identification_Information
type
Data_Quality_Information
type
Spatial_Data_Organization_Infor...
type
Spatial_Reference_Information
type
Entity_and_Attribute_Information
type
Distribution_Information
type
0..∞
Metadata_Reference_Information
type

Figure 3: BioMe Metadata schema – Metadata_BioME node –

4. INFRASTRUCTURE FOR METADATA AND DATA MANAGEMENT

Aiming to provide users access to biological data and metadata information, we have implemented a computational infrastructure mainly based on Open Source Technology (OST), where data broker and providers can store, modify and disseminate data and metadata through Internet. OST allows the development of powerful and fast database-generated web applications. The systems selected represent a compromise between software functionalities and costs. The objective was to provide a technical solution with a low cost, matching the requirement imposed at some institutions in the Amazon, that is, lack of funds for such initiatives. The solution comprises of a Client/Server architecture, where the interactions are performed through the Apache[2] Server. At the server side, we implemented the BioME Components, XML Server and Database Server Services.

BIOME XML COMPONENTS

The XML metadata BioME schema implements the FGDC standard, which includes the biological profile. Since the XML schema can ensure a well-formed, XML corresponding document, a template was developed, that is the XML file for BioME. The Metadata XML Documents hold descriptive material about the metadata schema, the html version and FGDC full documentation. Additionally, we made available the

[2] Apache is a product of the Apache Software Foundation; which provides support for the Apache community of open-source software projects.

XMLmind XML Editor (XXE[3]), an editor featuring a word processor-like view. It has been developed to make users comfortable and productive at editing XML documents and XML data. Users at a client environment can download XXE together with the Metadata XML template. Allowing the deployment of the XXE tool to users was the best option since Internet connections in the Amazon region present limitations in performance, and also the task to produce a metadata is time consuming. The template can be loaded into XXE and users can insert information that describes their metadata. All documents available for download can also be accessed for browsing via the Web interface. Once the metadata has been concluded, it is required that the metadata be registered and posted to the BioME site.

XML SERVER

We have tested two systems, Tamino[4] XML Server and XYZFind[5] Server. Even though the systems overlap in most of their functionalities, XYZFind was adopted due to its more simplified structure. XYZFind is an XML server that consists of an XML repository and an XML query engine. It accepts any number of a well-formed XML files and maintains a single data representation of the entire files stored. The XML files can be retrieved, updated, or removed from the catalogue via XYZQL requests. Once the repository has been indexed, functions of search and query are made available. The queries are written in XYZQL, providing support for path-level queries, Boolean queries, keyword search, and numeric range queries.

The XYZFind accepts an incoming HTTP request, processes the contained instructions, and issues an HTTP response. Any resource that can make an HTTP connection can be used to write an application that takes advantage of the server functions.

DATABASE SERVER

CLOSi schema descriptions (biological collections object classes, relationships) can be mapped into a relational database management system. We selected MySQL[6], because it is a fast, multi-threaded, multi-user and robust SQL database server.

The implemented architecture is presented in Figure 4.

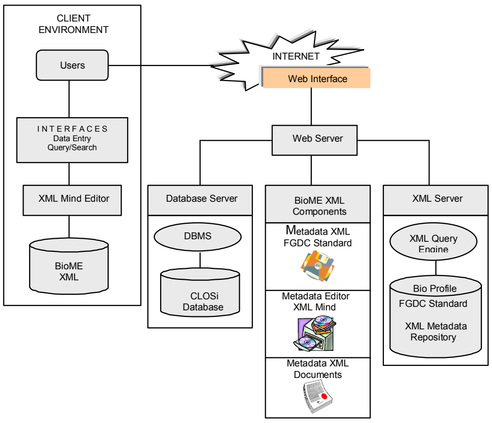

Figure 4: Client/Server architecture for data and metadata management

5. CONCLUSIONS

The present infrastructure has been developed and installed at ITC and is under test using data from INPA's Fish and Crustaceous collection. Researchers at INPA have started to produce metadata based on the XML template and editing it using the XMLmind Editor, both components deployed from the BioME XML Component site. A CLOSi schema of an entomological collection has been tested CLOSi partially. The results so far are encouraging and the Crustaceous data set has already started to be designed for its full extension. The current database implementation does not provide facilities to query via the web yet, as it is provided when querying the Metadata repository. The next step will be to provide XSL stylesheets for completed XML metadata files in order to render HTML pages over the web. In case of multiple output formats several XSL can be designed, that is, one stylesheet can render to HTML Navigator, Explore, Lynx, and to PDF for printing.

REFERENCES

Campos dos Santos, J.L., de By, R.A. & Magalhães, C. 2000. A case study of INPA's bio-DB and an approach to provide an open analytical database environment. International Archives of Photogrammetry and Remote Sensing, 33 (B4): 155-163.

Chen, I. A. & Markowitz, V. M. 1996. The Object-Protocol Model, Technical Report LBNL 32738, Lawrence Berkeley National Laboratory, Information and Computing Science Division, 1 Cycloton Road, Berkeley, CA 94720, June 1996.

DuBois, P. 2000. MySQL. New Riders Publishing, New York.

FGDC. 1998. Content Standard for Digital Geospatial Metadata. Federal Geographic Data Committee. Washington, D.C. pg. 85.

FGDC. 2001. Biological Data Profile Workbook. Federal Geographic Data Committee. Washington, D.C. pg. 148.

Graham, I.S. 1996. HTML Source Book. John Wiley & Son, London etc.

Sonderegger, J., Petry, P., Campos dos Santos, J.L. & Alves, N.F. 1998. An entomological collections database for INPA. In: Ling, T.W.; Ram, S. & Lee, M.L. (eds.), Proceedings of the 17th International Conference on Conceptual Modeling - ER '98. Singapore. p. 421-434.

[3] XXE Version 1.0 (October 2001) is a Pixware SARL product – Montigny Le Bretonneux, France.

[4] Tamino is a trademark of Software AG – Darmstadt, Germany.

[5] XYZFind is a trademark of XYZFind Corporation – Kirkland, WA, USA.

[6] MySQLis a trademark of MySQL AB. It can be used under the GNU General Public License (DuBois, 2000).