

A Biodiversity Semantic Associative Annotation Tool

David A. Gaitros
and Wei Zhang

Department of Computer Science
Florida State University
gaitrosd@cs.fsu.edu
wzhang@cs.fsu.edu

Greg Riccardi
College of Information
Florida State University
riccardi@ci.fsu.edu

Austin Mast

Department of Biological Sciences
Florida State University
amast@bio.fsu.edu

Fredrik Ronquist

School of Computational Science
Florida State University
ronquist@csit.fsu.edu

Abstract

MorphBank, an on-line museum quality collection of biological images, is an NSF funded project designed to facilitate the on-line collaboration of biologists from around the world [3]. This paper presents a methodology for the creation of an annotation tool for on-line collaboration and sharing of heterogeneous data through a common medium. The annotation tool combines the advantages of highly organized relational database, extensible XML schemas, Life Science Identifiers, and accepted industry ontologies using DataGrid technologies to facilitate collection and sharing information on biological specimens. Schematized annotation provides biologists with a flexible framework to perform annotations using their own data models. Structured XML documents enable structure-based semantic retrieval to improve the query accuracy. Retrieval performance can also be improved by combining the relational database and XML documents.

Index-Terms: bioinformatics, database, web services

1. Introduction

The discovery, identification, and documentation of biological entities is a time consuming and tedious task. The subtle differences between similar species may be so minute as to require the collaboration of several experts to identify. For any taxonomic group, there exist a number of such experts located around the world who can assist in the identification of specific organisms. Collaboration in species identification often involved the need for scientists to travel to the location of the specimens or for specimens to be sent to

the scientists for first hand examination.

There were several problems associated with the discovery of information in MorphBank. The first is finding an image associated with a specific species and genus, second is finding information about that image and it's association with other images, and third is the discovery of ad-hoc data about the images entered by biologists. Discovering ad-hoc data is the most problematic. As long as data is well formatted and constrained to the database schema then finding and retrieving it is simple. However, as we've discovered, there is no practical limit to the amount of information a scientist may wish to store with a particular specimen. Most of the knowledge is contained in the memory of these scientists or in hand written notebooks. Although it is recognized that manual annotation is expensive and time consuming it is nevertheless still essential in documenting collaborative knowledge in biological systems [2]. Translating and storing this knowledge in a searchable form is the challenge.

In this article, we present a method using existing technologies that allow scientists to use their own schemas in describing and storing information. Our approach combines the advantages of highly organized relational database and an extensible XML schema to provide a flexible framework and to offer an efficient semantic query. The remainder of this paper is organized as follows. Section 2 will provide the reader with a short background on the history of the MorphBank project. Then we will discuss biological annotation requirements in section 3. Section 4 describes the associative semantics and image annotations. Section 5 describes our preliminary tests using herbarium annotations.

2. Background

MorphBank is an open web repository of images serving the biological research community. MorphBank can serve as a virtual reference collection of named organisms or a resource for comparative morphological study; new use cases are continuously added [7]. Each image in the database is associated with fully search-able set of text information. Additionally images can be downloaded in several different formats [3].

The original MorphBank Database schema did not incorporate any ontology standards currently recognized in the industry [14]. The MorphBank Research team selected the Darwin Core set [9] for data item names and type representation. Although most biological ontology standards are relatively new, it was felt the Darwin Core set was the most complete. Using existing ontologies and other standards, MorphBank is desinged to be accessed through web services.

2.1. MorphBank Base Object Service

The MorphBank Base Object is a super class inherited by all MorphBank database objects. Each object contains a Life Science Identifier (LSID) [13] index. LSIDs assist us in cataloging the location of each object (and which service), the identification of the user who added the object, the date and time of creation, an optional description of the object, and the last time the object was modified. This feature allows anyone accessing MorphBank sufficient information to find and catalog data and associate related objects without implementing a varied number of keys with different data types. This service was a key element in the design of the new MorphBank system that makes annotations more meaningful and easier to implement. .

2.2. MorphBank Object Relationships

Since each MorphBank object is identified using LSIDs, the use of foreign keys within the database are not restricted to a single table. A single column within a MorphBank table holding an LSID may point to several different tables or a whole collection of objects. For instance, an Annotation object may be associated with an image, specimen, location, user, group, or even another complete set of annotations. This allows for the creation of complex collections of objects that can be shared with other users of the MorphBank system.

3. Biological Annotation Requirements

The users of the MorphBank database system have identified several requirements for image and object annotation

to be used by authorized users of the system. These requirements are in-line with the *Specifications for Image Annotation on the Semantic Web* as described W3C in their draft document [5]. A major restriction placed on the development was that the annotation software must be completely accessible through the use of a web browser. Additionally, annotations must be made in real-time and directly to the actual data source. Updates and annotations made by one scientist must be readily available to other colleges for collaboration in a timely manner.

There has been considerable effort put into the development of a general purpose web-based annotation tool sets over the past several years. In their paper on web annotations, Venu Vasudevan and Mark Palmer [15] described an approach 6 years ago on the development of a web based annotation tool that could be used to annotate documents over the Internet with just the use of a web browser. However, they discovered several limitations in the use of web browsers and of HTML as layout languages that made digital annotations somewhat cumbersome. The increase use of Java script, higher speed communications, improved web interface standards, and increased browser capability have made web based digital annotations more of a reality. However, there is still no convenient method for making annotations on the sides of web pages as you would on paper documents [8].

The problem of biodiversity annotation is simple. Biologists have increased the number of specimens they can gather but have not increased their ability to catalog, identify, and study them. Collaborations still include the exchange of physical specimens and the manual annotations of the images using indexed cards and paper documents. At the functional level, many users have developed their own specific but proprietary solution to this problem. Through the use of MorphBank and a web based annotation tool, we can solve most if not all of these problems.

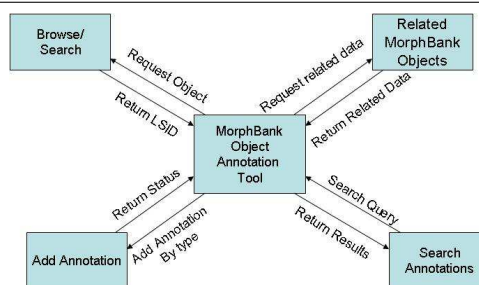


Figure 1. MorphBank Annotation Architecture

3.1. MorphBank Object Annotation

Current research involved with the development of annotation middle-ware products are currently focused on the development of automated laboratory notebooks such as those under development at the United States Department of Energy, National Collaboratories under the guidance of Dr. Jim Myers [11]. *These middle-ware products present researchers, applications, problem-solving environments (PSE), and software agents with a layered set of application services that provide a finite set of capabilities for the creation and management of meta-data, the definition of semantic relationships between data objects, and the development of electronic research records* [10]. However, many database applications developed using an evolutionary approach to design require the development of wrapper software in order to take advantage of these products [4]. These wrapper products translate the ontology of the database into a common definition using W3C standards such as an RDF Schema approach. Although an RDF Schema is very useful in some cases, the exact meaning of the data and related objects can be lost during the translation. MorphBank was designed to allow users to take advantage of web service products to gain access to the data by conforming to industry practices and standards but maintain the ontology of the original data. Figure 1 depicts the MorphBank Annotation Architecture as it is currently implemented. The database structure, table names, and data items are will documented and published. Users will browse or search the web site for MorphBank objects using a variety of tools provided through the web site.

3.2. Basic Annotation Template

Annotations are usually considered to be text associated with other text or images. Although the basic idea of annotations is rather old most web materials don't allow for annotations [6]. For annotation and search purposes, the MorphBank object annotation tool provides a minimum set of tools common to all annotation requirements. The tool uses the definitions as stated in the Darwin Core [1] initiative. The idea was to keep the tool-set as simple and as straight forward as possible. This was particularly important since all annotations must be made using only a web browser. The template for the tool defines several functional areas required for basic biodiversity annotation and specimen determination. Each functional area will map directly to an option on the MorphBank tool as a drop down menu item. The groups and individual menu items are described below:

- **Collection:** It is important that biologists be able to organize and work with various collections of images and specimens. Typically, scientists are working with a specific set of images from a single genus and species.

However, the number of images could number in the hundreds or thousands. For these reason, we have provided the capability for scientists to create a collection of images or related objects. Users can specify if they want to see a related set of images or objects grouped by specimen, view, taxonomic name, or even a private collection.

- **Mail:** Users may wish to send a link of an image to another user via email.
- **View:** Each MorphBank image has associated with it a pre-defined set of views where users can see different information about the image. Users can view a more complete set of data in the image, specimen, taxonomic structure, image view, and related publications by selecting the appropriate option.
- **Image Processing:** Biologists that were interviewed during the course of our research indicated a desire to temporarily manipulate images during the annotation phase to assist them in curation of the specimens. Biologists can adjust the image by rotating the object, zoom-in, zoom-out, or measuring specific portions using a set scale.
- **Annotation** This function deals with all of the aspects of organizing and adding annotations to the image currently displayed. Users can add a new annotation to the displayed image or sort the list of annotations by title, author, or date of annotation.

3.3. Type of Annotations

Using the ability to store XML documents with annotations gives us the ability to define associative semantic relationships with ad-hoc data and other MorphBank data. We have defined all existing and common annotations and then allow the user to define their own set of relationships and meanings. The categories of annotations are as follows:

- **General:** There are instances where users desire to make some ad-hoc comments concerning an image, specimen or other object in the database. The requirement for this type of annotation was made to allow maximum flexibility for including comments, measurements, and other related data to be stored and associated with the MorphBank Object.
- **Image:**As a phylogenetic database, images are vitally important to the users of the system. Therefore, many of the annotation types described in this section will apply specifically to images. The types of image annotations are listed as:
 - Spot location on an image associated with the annotation. The user will identify a specific spot on

the image to associate with a label, title, and paragraph description.

- Circle associated with an area on the image. The user will place a circle encapsulating an area to associate with a label, title, and paragraph description.
- Rectangle associated with an area on image. The user will place a rectangle encapsulating an area to associate with a label, title, and paragraph description.
- **Taxon Identification:** Used for discussion concerning the determination of a specimen. Users will select a specimen and by using the associated images, make a recommendation as to the specific genus and species determination.
- **Phylogenetic Character and State:** This type of annotation will be used to associate a phylogenetic character and state with a specific image or even a particular location on an image. In this type of annotation, the user will select from the database a genre of phylogenetic characters and a particular character and state.
- **Relationship:** MorphBank comes standard with pre-defined data relationships. Relationship annotations allow the user to define additional relationships associating MorphBank objects with each other. User will select any two MorphBank objects (image, specimen, view, location, publication, user, group, etc) and then describe the relationship among the two.
- **Schematized-User Defined:** MorphBank stores pre-defined XML schemas that define semantic associations between named data items that are not part of the static database. Users select one of these schemas and fill in the pertinent information. New schemas can be added at any time. This allows user to efficiently create complex general annotations that (on the surface) appear to be ad-hoc. This feature also has the added benefits of decreasing the search time for annotations and reliability of the information.

4. Associative Semantic Annotations

Specimen image annotation captures people's knowledge of species such as new observations, and disagreements with previous annotations. Image annotation enables semantic image retrieval and maintains a record of user comments concerning the data. Further more, a collection of featured annotations provides a way to assign species to a specimen. Image annotation associates textual information to the specific region of an image to enable semantic querying. Two technologies are frequently used: Text-based approach and field-based approach. The former simply add



Figure 2. A Simple Image Annotation Example

keywords to the whole image using natural language. However, keyword-based retrieval returns irrelevant documents (i.e., low accuracy of retrieval). Field-based method describes and retrieves an item using one or more field-value pairs, thus improves the retrieval precision. Figure 2 shows a simple image annotation of the field-based approach. However, both text-based and field-based approaches store the information in a plain text format. It is known that querying the plain text is inefficient. Furthermore, storing annotation information using only plain text is not suitable to satisfy the higher level requirements for the system. Meaning and ontology must be associated with the data. The heterogeneous data models from different biologists and the diversity of bioinformatics require frequent update and different data structure to the system. Dynamic creation tables in relational database for different data models is not a practical solution for MorphBank while taking integration constraints into consideration.

MorphBank stores the annotation information in a column of relational database using W3C XML document [16]. This structured, human readable and self-description XML document can be parsed using standard XML parser and can be easily extended. It has several advantages using the XML format. Firstly, it provides the biologists with a great deal of flexibility in annotating an image using their own data model. Secondly, it improves retrieval performance by combining relational database technology and XML querying such as XQuery [19] or XPath [18]. Thirdly, XML document can be easily validated by W3C XML schema [17].

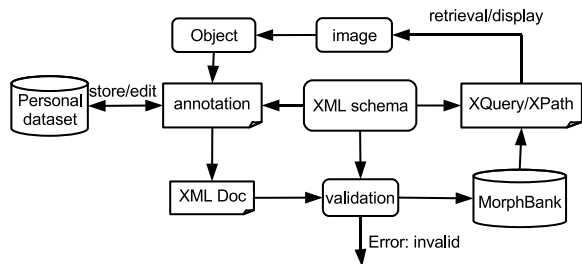


Figure 3. Image Annotation Overview

4.1. Associative Semantic Annotation Architecture

Figure 3 depicts the architecture of the image annotation. A user can browse or perform an XML query to select an image to annotate. An object is associated with a specific part of an image such as a box or a circle. Annotation interface is dynamically created based on the specific XML schema. User can choose a schema from a set of predefined schemas or create a new schema. Annotation can be saved in progress to be edited further, or can be imported into the MorphBank system. The information of annotation is stored in an XML document, an instance of the XML schema, which imposes structural and semantic constraints on it. Well-formedness verification and semantic validation of the XML document are performed against the schema. Structural or semantic errors indicate violation of the constraints and are returned to the user. Only valid XML documents are imported to the MorphBank system. Performance can be improved because no further verification or validation is required. Query interface is also generated dynamically from a specified schema. This specialized query improves retrieval performance by querying only the subset of the XML document.

4.2. Schema Customization

Annotations are usually considered to be text associated with other text or images. Although the basic idea of an annotation is rather old most web materials don't allow for annotations [6]. Because of the awkward nature of most web-based annotation tools, they are seldom used and as a result most scientists use manual annotations in their day-to-day research. The challenge in a web-based annotation tool is providing enough functionality given the limitations of most web browsers to make them useful to researchers. User defined schemas extend the predefined schemas to meet individuals needs. All the XML schemas are derived from a generalized schema class, which defines the general information of an image. This general schema is listed be-

low¹.

```

<schema>
  <simpleType name="id-type">
    <restriction base="xs:string">
      <pattern value="[1-9][0-9]{4}"/>
    </restriction>
  </xs:simpleType>
  <simpleType name="name-type">
    <restriction base="xs:string">
      <maxLength value="255"/>
    </restriction>
  </xs:simpleType>
  <complexType name="creator-type">
    <sequence>
      <element name="lastname" type="name-type"/>
      <element name="firstname" type="name-type"/>
      <element name="title" type="name-type"/>
    </sequence>
  </complexType>
  <complexType name="point-type">
    <sequence>
      <element name="x" type="xs:integer"/>
      <element name="y" type="xs:integer"/>
    </sequence>
  </complexType>
  <complexType name="rec-type">
    The use of LSIDs allows MorphBank to
    underlying database structure and permits the replacement of internal
    items with external references
    <sequence>
      <element name="point" type="point-type"
        minOccurs="2" maxOccurs="2"/>
    </sequence>
  </complexType>
  <!-- other declaration such as circle, polygon -->
  <complexType name="loc-type">
    <choice>
      <element name="rectangle" type="rec-type"/>
      <!-- other loc types such as circle, polygon -->
    </choice>
  </complexType>
  <complexType name="obj-type">
    <sequence>
      <element name="name" type="name-type"/>a
      <element name="location" type="loc-type"/>
      <element name="description" type="xs:string"/>
    </sequence>
  </complexType>
  <complexType name="annotation-type">
    <sequence>
      <element name="image-id" type="id-type"/>
      <element name="LSID" type="id-type"/>
      <element name="object" type="obj-type"/>
      <element name="curator" type="curator-type"/>
      <element name="date" type="date-type"/>
    </sequence>
    <attribute name="id" type="id-type">
    <attribute name="type" type="name-type">
  </complexType>
  <!-- top level element -->
  <element name="annotation" type="annotation-type">
</schema>
  
```

This schema defines a set of general information such as annotation-id, annotation type, image-id, curator and created date. All the image annotations contain those general information. For example, Figure 4 shows an XML document of the general schema. The annotation consists of a rectangle region specified by a sequence of coordinates on the image representing the top left corner and right bottom corner respectively.

¹ slightly simplified for clarity and size.

```

<annotation id = '12345' type='new'>
  <imageid>34567</imageid>
  <LSID>11223</LSID>
  <object>
    <name>leaf</name>
    <location>
      <rectangle>
        <point><x>100</x><y>200</y></point>
        <point><x>300</x><y>400</y></point>
      </rectangle>
    </location>
    <description>
      This is a piece of leaf ...
    </description>
  </object>
  <creator>
    <lastname>Darwin</lastname>
    <firstname>Darwin</firstname>
    <title>Biologist</title>
  </creator>
  <date>08-26-2005 10:28:36</date>
</annotation>

```

Figure 4. An Example XML Annotation Document for an Image

4.3. Annotation and Schema Interfaces

A graphic annotation interface is automatically and dynamically generated for users from a specific schema. An image is displayed with highlighted region and annotation information. A set of text fields based on the schema structure are created. Some schema defined data types such as enumeration are created for users as a list of choices. Users can save their works or submit the annotations to the system. XML documents are automatically generated and validated against the schema. A schema interface is also created if a user chooses to create a new schema. User-defined schemas are uploaded and stored in the system.

4.4. Query Interface

Similarly as an annotation interface, a graphic query interface is created from a specific schema. The schema-specific and structure-based retrieval improves the accuracy for semantic query. Since each XML document stored in a relational database column and indexed by the schema, only the sub set of XML documents that are indexed by the schema are searched. In addition, MySQL 5.1 provides native XML functions for searching and changing XML documents [12]. The query interface automatically generates the optimized querying to improve the query performance.

5. Preliminary Results

The MorphBank research team has been working closely with a group of botanist at the Department of Biological Sciences at Florida State University to use the *beta*

Figure 5. Sample - Herbarium Annotation

version of the annotation tool for the curation of herbarium specimens. The team took a standard herbarium annotation card (see Figure 5) and created an XML schema which is stored in the MorphBank database as a text field in the Annotation Schema table and identified with an LSID. The annotation software, when displaying annotations, can determine if the text of the annotation is either plain ascii or an XML document and select the correct application to display the data. Web services are used to verify if an XML document is valid and conforms to the stored schema.

Users making taxonomic identification annotations on herbarium collections would then select the schema associated with this type. The XML schema mirrors the information and organization of the original annotation card. The software requests the user to fill in the missing information and stores the data as an XML document as shown in Figure 6. The results were very promising. The ability to expand the meaning of annotations increased the utility of the software to the point where scientists were able to use it for species determination. User of MorphBank were able to store their data into MorphBank using their own datasets. Additional information they wish to include was placed in the system as user-defined annotations. The user extracts their schema from their personal database as an XML document and submits the document to MorphBank to be included as a validated annotation schema. Once accepted into the database, scientists can import their data directly into MorphBank without altering the baseline schema. They simply selected an image and requested to add an annotation. Since the data is in an XML format it is highly search able using existing Web services tools. Additionally, the native ontology of the dataset is preserved avoiding the problem of translation of the meaning of the data.

```

<annotation id = '12345' type='Taxon Identification - Herbarium'
  <imageid>34567</imageid>
  <LSID>11223</LSID>
  <object>
    <name>Psilocybe Dumontii</name>
    <location>
      <rectangle>
        <point><x>100</x><y>200</y></point>
        <point><x>300</x><y>400</y></point>
      </rectangle>
    </location>
    <description>
      This is apparently an undetermined specimen
      of polydesmia. Probably close to
      Psilocybe Dumontii.
    </description>
  </object>
  <creator>
    <lastname>Koff</lastname>
    <firstname>Richard I. </firstname>
    <title>Biologist</title>
  </creator>
  <date>08-26-2005 10:28:36</date>
</annotation>

```

Figure 6. Herbarium Taxonomic Identification

6. Conclusion

We have described an existing need in the biological community to store and retrieve complex information on specimen and related images. In creating a web site that stores the elements common to all entities in the tree-of-life, we limit our ability to store data that is unique to specific species. Through the use of an extensible schema and associating semantics through an XML document stored as an annotation, we increase the flexibility and utility of such a system.

Our work in developing a web-based annotation tool has proven successful. The non-intrusive method, permits biologists to mark images without altering the original image, and share this annotations with others in an easy and open format. Often, mapping information into an abstract form requires developers and designers to alter the structure of real world relationships in order to fit a specific paradigm. We hope to have shown in this paper a method that will allow users of MorphBank to both have the integrity of a well designed centralized digital image database with the flexibility of a privately owned collection. Our hope is that the work performed under this NSF grant by the MorphBank project will provide the Tree-of-Life initiative with a stable digital image database and annotation tool set that can be used by biologists around the world.

References

- [1] L. Alexander, A. Runyan, and V. Anderson. Taxonomic data working group, darwin core 2.
- [2] A Dingli, F Ciravegna, and Y Wilks. Automatic semantic annotation using unsupervised information extraction and inte-

- gration. In *Workshop on Knowledge Markup and Semantic Annotation, KCAP03*, 2003.
- [3] D. Gaitros, G. Riccardi, F. Ronquist, N. Jammigumpula, and W. Blanco. Morphbank, the development of a general purpose bioinformatics database. *Conference on Internet Computing (ICOMP'05)*, pages 31–37, Jun 2005.
- [4] L. Haas, D. Kossmann, E. Wimmers, and J. Yang. An optimizer for heterogeneous systems with non-standard data search capabilities. in special issue on query processing for non-standard data. *IEEE Data Engineering Bulletin 19(4)*, pages 37–43, Dec 1996.
- [5] C Halasheck-Weiner, J Hunter, N Simou, J Smith, and V Tzouvaras. Image annotation on the semantic web, Jan 2006.
- [6] P. Korica, H. Maurer, and N. Scerbakov. Extending annotations to make the truly valuable. *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (ELEAN) 2005*, 2005.
- [7] J Liljeblad and F Ronquist. A phylogenetic analysis of higher-level gall wasp relationships (hymenoptera: Cynipidae). *Systematic Entomology*, 23:229–252, 1998.
- [8] P. Marshall. Annotations: From paper books to the digital library. in *Proceedings of the ACM Digital Libraries 97 Conference, Philadelphia, Pa*, Jul 1997.
- [9] C Meng. Biological information standards. *Bulletin of the American Society for Information Science and Technology*, 2004.
- [10] j Myers. <http://collaboratory.emsl.pnl.gov/>, 2004.
- [11] J Myers, A Chappell, M Elder, A Geist, and Schwidder J. Re-integrating the research record. *IEEE Computing and Science and Engineering*, May 2003.
- [12] MySQL. <http://dev.mysql.com/tech-resources/articles/mysql-5.1-xml.html>.
- [13] D. Smith S. Martin and B. Szekely. Lsid(life science identifier)project, 2005. <http://lsid.sourceforge.net>.
- [14] P Spyns, R Meersman, and M Jarrar. Data modeling versus ontology engineering. *SIGMOD Record*, 31(4):12–17, December 2002.
- [15] V. Vasudevan and M. Palmer. On web annotations: Promises and pitfalls of current web infrastructure. *32nd Hawaii International Conference on Systems Sciences*, Jan 1999.
- [16] W3C. <http://www.w3.org/XML/>.
- [17] W3C. <http://www.w3.org/XML/Schema>.
- [18] XPath. <http://www.w3.org/TR/xpath>.
- [19] XQuery. <http://www.xquery.com/>.