

ADDING VALUE TO BIODIVERSITY IMAGES THROUGH COMMUNITY ANNOTATION

Greg Riccardi
College of Information
Florida State University
Tallahassee, FL 32306-2100 USA
01-850-644-2869
Riccardi@ci.fsu.edu

Andrew Deans, David Gaitros, Katja Seltman, Steven Winner
Neelima Jammingumpula, School of Computational Science
Corinne Jorgensen, Peter College of Information
Jorgensen, Austin Mast, and Department of Biological Science
Karolina Maneva-Jakimoska, Florida State University
Debbie Paul, Fredrik Ronquist,

ABSTRACT

Morphbank, an on-line collection of museum-quality biological images, is an NSF funded project designed to facilitate the on-line collaboration of biologists from around the world [3]. Our primary focus is to aid in the collection and management of images that are useful in phylogenetic research. Morphbank users are actively collaborating on the creation of information that represents the associations among images and related biodiversity data objects. This paper describes the Morphbank annotation tool and data models and gives examples of how users create structured information in the system. Schematized annotation provides biologists with a flexible framework to create semantically-rich annotations using their own data models.

Keywords

Annotation, association, biodiversity

1. INTRODUCTION

The discovery, identification, and documentation of biological entities are time consuming and tedious tasks. The subtle differences between similar species may be so minute as to require the collaboration of several experts to identify. Each taxonomic group has many experts who can assist in the identification of specific organisms. However, with the increase in the number of new organisms that have been discovered and a decrease in number of senior specialists, identification and curation of data have become more difficult. Often, it 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. This is still standard practice among most biologists today.

Morphbank contains information about organisms. Each image in the system is associated with one or more specimens. Each specimen is a representation of information about an organism. Specimens are in turn associated with localities, contributors, taxonomic concepts, and a variety of annotations.

The design and development of the Morphbank system identified several challenges in discovering and creating information about images and their related objects.

- Finding images and specimens associated with a specific species and genus,
- Finding and recording information about that image and its related objects, and
- The discovery and recording of ad-hoc associations among the various objects.

Discovering and recording ad-hoc data is the most problematic. It is particularly difficult to find ways that users can record associations among objects.

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.

2. BACKGROUND

Morphbank is an open Web repository of images serving the biological research community. It is currently being used to document specimens in natural history collections, to voucher DNA sequence data, and to share research results in disciplines such as taxonomy, morphometrics, comparative anatomy, and phylogenetics. 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 searchable set of text information. Additionally images can be downloaded in several different formats [3]. Understanding the background of Morphbank is important to understanding the complexity of the problem of collaborating with other scientists on the identification and curation of biodiversity data.

2.1 MORPHBANK OBJECTS

Each object in the Morphbank system is uniquely identified and includes a set of standard fields that assist us in cataloging the location and type of each object, 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. These attributes allow anyone accessing Morphbank sufficient information to find and catalog data and associate related objects. Each object is externally identified by a Life Science Identifier (LSID) [13].

2.2 MORPHBANK OBJECT RELATIONSHIPS

Since each Morphbank object is uniquely identified, any object can be the target of a stored reference. A single column within a Morphbank table holding a foreign key may refer to several an object of any type. Thus a collection object can be heterogeneous. For instance, an annotation object may define an association among images, specimens, locations, users, or even other annotations.

¹ Supported by NSF contract DBI-0446224, 2005-2008
WWW 2007, May 8--12, 2007, Banff, Canada.

This flexibility allows for the creation of complex collections of objects that can be shared with other users of the Morphbank system. Although there are a series of predefined relationships in Morphbank, the use of unique identifiers allows users to define an unrestricted set of complex relationships of objects within the confines of the system.

Figure 1 shows the result of searching for images that are related to the taxon with id 30244, the species *Asclepias amplexicaulus*. The search looks through the known associations between objects to find the proper set. Each image in the set is associated with a specimen which is associated with the proper taxon. The structure of these predefined associations allow the search to be both effective and efficient. The information about the images in Figure 1 comes from the image, its related specimen and its related taxon.

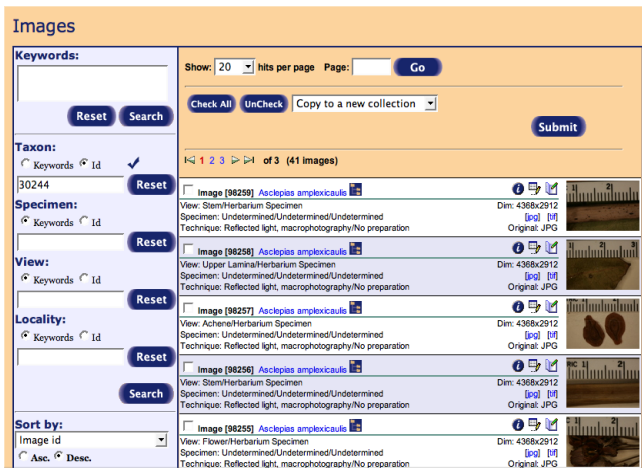


Figure 1. The result of searching for images for a particular taxon

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 consistent with the *Specifications For Image Annotation On The Semantic Web* as described W3C in their draft document [5]. A major restriction placed on Morphbank development was that the annotation software must be accessible through the use of a Web browser without the need to download an extensive set of client based applications. This requirement was established because research biologists frequently travel from one location to another and many times only have access to a Web browser. Additionally, annotations must be made in real-time and directly to the actual data source to avoid update anomalies associated with multiple copies of the data. 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 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 increased use of Javascript, 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 that 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.

3.1 MORPHBANK OBJECT ANNOTATION

A variety of annotation technologies allow users to add value to images by creating associations between those images, text and other digital objects. Morphbank takes this one step further by making the associations into first class objects that can themselves be annotated and associated with other objects. Morphbank also allows associations to take on specific semantic characteristics that constrain their meaning and thereby improve searching and understanding.

Image annotation is available in a variety of image management Web sites. The simplest annotations are found in systems that support attaching tags to images and other media. Flickr.com and YouTube.com, e.g., allow users to add text attributes (tags) to images and use those tags to support searching. FotoTagger.com, among others, goes a step further and allows the tags to be attached to specific locations on images.

Blogging is another form of image annotation in which text passages are linked to images, Web pages and other digital objects. A blog entry creates an associate between its own text and the linked objects.

Annotea.org supports the creation of RDF attributes for image tags. These attributes can be used to provide search inference capabilities for users of image repositories.

Another annotation strategy involves the development of 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]. Users are able to record associations between digital objects across and among projects.

Morphbank seeks to combine these ideas by allowing incorporating an extensible annotation type system and by systematically expanding the scope of associations by including any objects referenced by globally unique IDs (GUID).

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. 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

An annotation is an assertion that a collection of objects are related in a particular way. 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 terminology of the Darwin Core [1] biodiversity ontology initiative. We strove to keep the tool-set as simple and as straight forward as possible and to provide specializations that make it easy for particular types of annotations to be created.

Flexibility is particularly important because 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.

3.3 TYPES OF ANNOTATIONS

Using the ability to store complex metadata with annotations gives allows us to define associative semantic relationships with ad-hoc data and other Morphbank data. The data model that supports annotation is intended to be extended to incorporate additional types as needed by users. The categories of annotations in the current system are as follows:

- **General:** There are instances where users desire to make some ad-hoc comments concerning a collection of images, specimens or other objects. 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 collection of objects. A very useful example of a general annotation is a simple collection of objects, much like a shopping cart, that can be stored, organized, and labeled for later use.
- **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 Determination:** Used for discussion concerning the species or other taxonomic 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. Taxon determinations are extremely important to the research activities of the primary users.
- **Phylogenetic Character and State:** This type of annotation will be used to organize physical features (called “characters”) of organisms into objects of interest to research users. Phylogenetic characters and possible values (states) of those characters are associated with specific images, with species, and with collections of species. In this type of annotation, the user will associate an image or specimen in the database with phylogenetic characters and states.
- **Relationship:** Morphbank comes standard with predefined 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.

4. EXAMPLES OF 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. Furthermore, 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 keywords to the whole image using natural language. However, keyword-based retrieval returns irrelevant documents (i.e., low accuracy of retrieval). A field-based method describes and retrieves an item using one or more field-value pairs, thus improves the retrieval precision. Figure 2 shows an image annotation of the field-based approach. This annotations asserts that a particular portion of an image (of a wasp leg) is a femur.

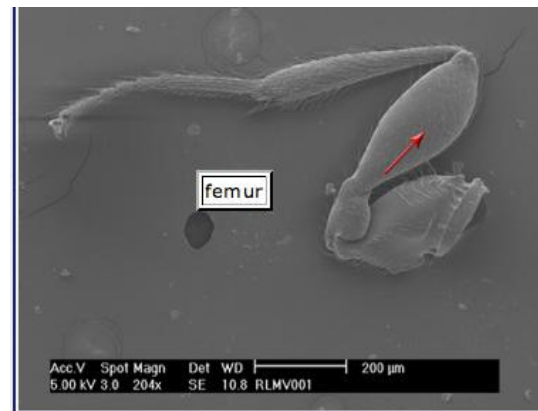


Figure 2. An Image Annotation Example

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 association types require frequent update and evolving data structures.

Figure 3 shows a Morphbank image annotation in context. The annotation contains attribution (upper left), a small instance of the annotated image (upper right), detailed comments, with technical terms highlighted (lower left), and brief descriptions of other annotations of the same image (lower right).

The annotation of Fig. 3 asserts that the wasp whose leg is shown has a particular feature, which is called “femur swollen medially”. Such features are used by experts to categorize specimens into taxonomic units (genus, species, etc.) and, after analysis, to develop evolutionary models.

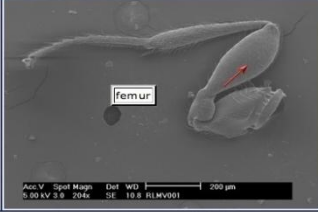
Morphbank is using annotation and association technology to collect information that is directly used in scientific research.

Each of the Morphbank objects related to the annotation of Figure 3—the image, the annotations, the related specimen, etc.—are represented as first-class objects with globally-unique identity. Thus the objects can be stored in collections, included in other annotations, and referenced in external sites.

Annotation Record: [109845] Title = femur

Contributed By: Andrew Deans
 Date Contributed: 10-07-2006
 Last Modified: 10-07-2006
 Publish Date: 04-06-2007

Specimen Id: [78506]
 Sex: Male
 Collector: B. L. Fisher et al.
 Species Name: *Ceraphron* sp.RLMV005
 Object Id: [78554]
 Object Type: Image
 Type of Annotation: General



Comments

Femur swollen medially; imbricate sculpturing throughout anterior and posterior sides; hairs present and evenly distributed

Related Annotations to this image

TITLE: unguis
 TYPE ANNOTATION: General
 BY: Andrew Deans
 DATE CREATED: 2006-10-07 16:04:44
 RELATED ANNOTATIONS OF ID: [109844]
 SINGLE SHOW OF ANNOTATION ID: [109844]

TITLE: femur
 TYPE ANNOTATION: General
 BY: Andrew Deans
 DATE CREATED: 2006-10-07 16:06:37
 RELATED ANNOTATIONS OF ID: [109845]
 SINGLE SHOW OF ANNOTATION ID: [109845]

Figure 3. Image Annotation In Context

Mass annotations are possible as well. Figure 4 shows an interface that allows a user to annotate each of a group of objects. In this case, the user is preparing to comment on the species identification, also called the *determination* of several botanical specimens. This annotation interface has been developed to enable a specific activity to be performed by experts on plant morphology.

10 Images of 10 in Collection 109267 [*Asclepias amplexicaulis* whole specimens]

Record: [19182] Record: [19183] Record: [19184] Record: [19185] Record: [19186] Record: [19187] Record: [19188] Record: [19189]

Type of Annotation * Determination

Related Annotations

Taxonomic Name	Taxon Author	Prefix	Suffix	A	D	S
<input type="radio"/> <i>Asclepias amplexicaulis</i>	Sm.	none	none	19	0	9
<input type="radio"/> <i>Asclepias amplexicaulis</i>	Sm.	?	none	1	0	1

Determination Annotation

Determination Data Fields

Determination Action *

New Taxon

Prefix

Suffix

Materials used in Id

Source of Identification *

Resources used in Identification *

Figure 4. Group Annotations

5. PRELIMINARY RESULTS

The Morphbank research team has been working closely with a group of botanists at the Department of Biological Sciences at Florida State University to use the annotation tool for the curation of specimens from the Robert K. Godfrey Herbarium at Florida State University. Figure 5 shows some of the Morphbank information for a typical herbarium sheet.

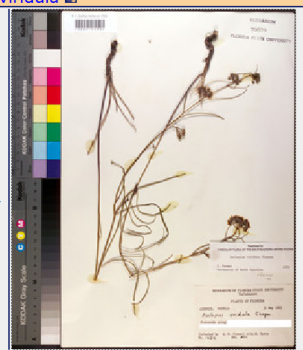
Image Record: [92372] *Asclepias viridula*

Contributor: Austin Mast
 Submitter: Debbie Paul
 Submitted date: 07-19-2006
 Published date: 07-15-2006

Access #: 44338
 Magnification: NULL
 Dimension (px): 3000x3582
 Resolution (PPI):
 Submitted as: JPG

View id: 77407
 Specimen part: [77359] - Plant body
 Angle: [77360] - Herbarium Specimen
 Technique: [27] - Reflected light, macrophotography
 Preparation: [77344] - No preparation

Download: tiff (30.96 MB)
 jpeg (1.08 MB)



Specimen

Specimen id: 91931
 Basis of record: [5] - Specimen
 Sex: [79416] - Undetermined
 Form: [77362] - Undetermined
 Stage: [79192] - Undetermined
 Collector: Andre F. Clewell, E. Tyson
 Date collected: 05-31-1963

Locality

Locality Id: 81529
 Continent ocean: [NA] - NORTH AMERICA
 Country: [US] - UNITED STATES
 Locality: Florida; Wakulla County
 Latitude:
 Longitude:
 Elevation (m):

Determination

Class: *Magnoliopsida*

Determination annotations

Figure 5. Morphbank display of the image of a herbarium sheet

Creating the determination annotation sheet began with interviews with domain experts and the evaluation of typical manual records. Figure 6 shows a detail of the herbarium sheet of Figure 5 that contains the information cards that are attached to the sheet. Two cards are attached. The lower card is the primary information about the specimen including who collected it, when and where. The lower card also shows the species determination that was recorded when the specimen was collected.

Examined for
 VASCULAR FLORA OF THE SOUTHEASTERN UNITED STATES

Asclepias viridula Chapman

J. Farmer
 University of North Carolina
 1983

V.D.D. WARD
 1985

HERBARIUM OF FLORIDA STATE UNIVERSITY
 Tallahassee

PLANTS OF FLORIDA

COUNTY: WAKULLA
 31 May 1963

Asclepias viridula Chapm.

Flatwoods along

Collected by A. F. Clewell with E. Tyson
 No. 1654
 Det. AFC

Figure 6. Information card from herbarium sheet

The upper card shows a determination annotation that was added to the specimen in 1983. J. Farmer of the University of North Carolina agreed that the determination was correct.

In pencil, between the two cards is second annotation. D. D. Ward in 1983 also agreed on the correctness of the determination.

The Morphbank annotation tool is intended to allow the online collection and dissemination of information like that shown in

Fig. 6. The tool will allow researchers to evaluate the determination of the specimen, that is, the association between each specimen and its taxon. The activity is an evaluation of the quality of the information stored in the herbarium.

A major benefit of the Web tools is its support for distributed collaboration. Before the sheets were

The annotation interface shown in Fig. 4 can be used to agree with the recorded determination of the set of specimens, or to disagree and select a different taxon. In this way the annotation represents a qualitative evaluation of the recorded information. Fig 4 shows that 19 annotations already record agreement (A) with the determination.

The results so far are very promising. Fifteen taxonomists were asked to use Morphbank images of specimens from the Robert K. Godfrey Herbarium at Florida State University to make digital determination annotations for 50 specimens each. The scientists found the online tools to be an excellent replacement for the manual task. They were particularly pleased to be able to see the results online and to be able to see the effects of this online collaboration.

An additional study of the feasibility of making determinations from images in lieu of physical specimens was conducted by bringing some of these experts to Florida. The study is ongoing. We hope to be able to establish that digital representations of these specimens are more than adequate replacements for the real objects.

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 have made biodiversity research more effective.

Our work in developing a tool that allows users to annotate images via the Web using only the essential elements 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. 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.

7. REFERENCES

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

integration. 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 Halaschek-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 Liljebblad 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 Eider, A Geist, and Schwidder J. Reintegrating the research record. IEEE Computing and Science and Engineering, May 2003.
- [12] MySQL. <http://dev.mysql.com/techresources/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. possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7”).