

**THE FLORIDA STATE UNIVERSITY
COLLEGE OF ARTS AND SCIENCES**

**THE REPRESENTATION OF ASSOCIATION SEMANTICS WITH
ANNOTATIONS IN A BIODIVERSITY INFORMATICS SYSTEM**

By

DAVID A. GAITROS

**A Dissertation submitted to the
Department of Computer Science
In partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy**

**Degree Awarded:
Spring Semester, 2007**

**Copyright © 2006
David A. Gaitros
All Rights Reserved**

The members of the Committee approve the Dissertation of David A. Gaitros
defended on December 8th, 2006.

Greg Riccardi
Professor Directing Dissertation

Fredrik Ronquist
Outside Committee Member

Robert van Engelen
Committee Member

Ashok Srinivasan
Committee Member

Approved:

David Whalley, Chair, Department of Computer Science

The Office of Graduate Studies has verified and approved the above named committee members.

This dissertation is dedicated to my wife Cynthia and our three children Heather, Adria, and Anthony. Without them, this work would have no meaning or purpose.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor and friend, Dr. Greg Riccardi for his guidance, support and patience all of these years. I would also like to thank Dr. Fredrik Ronquist and other members of the MorphBank research team for the opportunity to work with them in the area of biodiversity. Through the MorphBank grant we were able to accomplish something truly worthwhile. I would like to personally thank the MorphBank development team of Wilfredo Blanco, Neelima Jammigumpula, Steve Winner, Karolina Maneva-Jakimoska, Debra Paul, Katja Seltmann, and Cynthia Gaitros for their hard work and dedication. There has never been a more productive team. I owe a special thanks to the late Dr. Richard W. Hamming for his encouragement to pursue this degree.

Finally, I would like to thank my wife Cynthia for her love, encouragement, help, and support throughout the years. Without her, this degree would not have been possible. We always do our best work together.

TABLE OF CONTENTS

List of Tables	viii
List of Figures	ix
Abstract	xii
1. RESEARCH INTRODUCTION	1
1.1 Problem Definition.....	3
1.2 Research Goals	5
1.3 Annotation of Metadata Relationships	7
1.4 Verification of Research Objective.....	8
1.5 Research Accomplishments	8
1.6 Research Activities	9
1.7 Research Challenges	9
1.8 Chapter Summary	11
2. ANNOTATIONS IN LARGE SCIENTIFIC DATABASES	12
2.1 Automatic Semantic Annotation.....	12
2.2 Gene Sequence and DNA Annotation	14
2.3 Semantic Annotations in Image Collections.....	17
2.4 Challenges of Annotations on the Web	20
2.5 Scientific Annotation Middleware (SAM)	22
2.6 Chapter Summary	23
3. REQUIREMENTS OF A WEB BASED IMAGE AND PHYLOGENETIC DATABASE SYSTEM	24
3.1 The Motivation for MorphBank	24
3.2 MorphBank Security Requirements	27
3.3 Data Access Requirements	33
3.4 MorphBank Data Requirements	34
3.5 External Object Exposure	38
3.6 Rudimentary Query Requirements	41
3.7 Chapter Summary	45
4. MORPHBANK CONCEPTUAL MODEL	47
4.1 Base Object Relationship.....	48

4.2 Image and View Relationship.....	49
4.3 Specimen, Image, and Locality Relationship	50
4.4 Collection Relationships	51
4.5 Related Objects	53
4.6 Annotation Conceptual Model	54
4.7 Security Conceptual Model	58
4.8 Life Science Identifiers.....	61
4.9 Web Services	63
5. SEMANTIC ANNOTATIONS	68
5.1 MorphBank Base Object Service.....	71
5.2 Biological Annotation Requirement	71
5.3 Association Annotations	74
5.4 Image Annotations	77
5.5 Annotation Integrity.....	80
5.6 Preliminary Results.....	84
5.7 Conclusions.....	85
6. ANNOTATION TRIALS	86
6.1 Annotation Trial Objectives.....	86
6.2 Trial Procedures	87
6.3 Trial Participants.....	88
6.4 Initial Trial Feedback.....	88
6.5 Image and Specimen Data Summary	89
6.6 Collection Data Summary	91
6.7 Annotation Data Summary	92
6.8 Determination Annotation Data Summary	96
6.9 Chapter Summary	98
7. CONCLUSION	100
7.1 Research Results	101
7.2 Research Objectives Achieved	101
7.3 Annotation of Metadata Relationships	102
7.4 Verification of Research	102
7.5 Accomplishments of Research Challenges.....	108
7.6 Future Work.....	109

APPENDICES	110
A MorphBank Contributors.....	110
B Darwin Core Standard 2.0	112
C Version 2.5 Annotation User's Manual	118
D Example: MorphBank XML Image Annotation Schema	138
BIBLIOGRAPHY.....	140
BIOGRAPHICAL SKETCH	146

LIST OF TABLES

Table 2-1: Thesaurus Reference	19
Table 3-1: Relationship Symbols.....	41
Table 4-1: MorphBank Base Objects.....	47
Table 4-2: Object-to-Object Relationships	54
Table 5-1: Animal Kingdom Taxonomic Rank Id.....	83
Table 6-1: MorphBank Contributors	88

LIST OF FIGURES

Figure 1-1: Left Fore View, Female	3
Figure 2-1: Annotation of Gene Sequences	15
Figure 2-2: RDF Schema	17
Figure 2-3: “The Birthday” Painting of Chagall	18
Figure 3-1: TreeBase Search Console	26
Figure 3-2: MorphBank Architecture	33
Figure 3-3: MorphBank Object Model	35
Figure 3-4: LSID Examples	39
Figure 3-5: Sample RDF	40
Figure 3-6: Image Queries	42
Figure 3-7: Specimen Queries	42
Figure 3-8: View Queries	43
Figure 3-9: Locality Queries.....	43
Figure 3-10: User and Group Queries.....	44
Figure 3-11: Collection Queries	45
Figure 3-12: Annotation Queries	45
Figure 4-1: User and Object Relationships.....	48
Figure 4-2: Image and View Relationships	50
Figure 4-3: Specimen, Image and Locality Relationships.....	51
Figure 4-4: Collection Relationships	53

Figure 4-5: Multiple Relationship Diagram.....	55
Figure 4-6: Annotation Conceptual Model.....	56
Figure 4-7: Annotation Inheritance Model.....	57
Figure 4-8: Annotation Display Model	59
Figure 4-9: MorphBank User Privilege Use Cases.....	62
Figure 4-10: LSID MorphBank Examples.....	64
Figure 4-11: IBM Web Services Architecture.....	65
Figure 4-12: MorphBank Web Services Architecture	67
Figure 5-1: MorphBank Schema.....	69
Figure 5-2: MorphBank Annotation Architecture	70
Figure 5-3: Simple Image Annotation Example	76
Figure 5-4: Image Annotation Overview	77
Figure 5-5: An Example XML Annotation Document	79
Figure 5-6: Sample – Herbarium Annotation	80
Figure 5-7: Herbarium Taxonomic Determination.....	84
Figure 6-1: Specimen Single Show Example	90
Figure 6-2: Collection Show Example.....	91
Figure 6-3: Sample Single Image Annotation	95
Figure 6-4: Sample Determination Annotation	97
Figure 6-5: Sample Mass Determination Annotation	99
Figure 7-1: Sample Text Search	106

Figure 7-2: XML Search Screen.....	107
Figure 7-3: XML Search Results Screen	107

ABSTRACT

A specialized variation of associations for biodiversity data is defined and developed that makes the capture and discovery of information about biological images easier and more efficient. *Biodiversity* is the study of the diversity of plants and animals within a given region. Storing, understanding, and retrieving biodiversity data is a complex problem. Biodiversity experts disagree on the structure and the basic ontologies. Much of the knowledge on this subject is contained in private collections, paper notebooks, and the minds of biologists. Collaboration among scientists is still problematic because of the logistics involved in sharing collections. This research adds value to image repositories by collecting and publishing semantically rich user specified associations among images and other objects.

Current database and annotation techniques rely on structured data sets and ontologies to make storing, associating, and retrieving data efficient and reliable. A problem with biodiversity data is that the information is usually stored as ad-hoc text associated with non-standardized schemas and ontologies. This research developed a method that allows the storage of ad-hoc semantic associations through a complex relationship of working sets, phylogenetic character states, and image annotations. MorphBank is a collaborative research project supported by an NSF BDI grant (0446224 - \$2,249,530.00) titled “*Web Image Database Technology for Comparative Morphology and Biodiversity Research*”. MorphBank is an on-line museum-quality collection of biological images that facilitates the collaboration of biologists from around the world. This research proves the viability of using association semantics through annotations of biodiversity informatics for storing and discovery of new information.

CHAPTER 1

RESEARCH INTRODUCTION

I show in this dissertation that informal information contained in images, physical specimens, hand written laboratory notes, and the memory of scientists can be formally captured in an environment that provides a forum for collaboration using semantically rich associations. Scientific research conducted in the areas of biology, chemistry, meteorology, physics, and other disciplines rely on a scientist's intimate knowledge of both formal and informal data repositories. The discovery of information relies on the ability of the scientists to access the correct data sources. Distributed databases have emerged as a means of sharing large amounts of information among collaborating organizations [CFKST01]. Informal catalogs of unformatted data can be organized into a more systematic form. I developed examples of tools in MorphBank that implement the theory of a more sophisticated information discovery and retrieval method.

This research involved the following activities:

- a) I developed heuristics that ensures metadata can be captured accurately as possible.
- b) I examined examples of ad-hoc legacy annotations and hand written notes on biology specimens and analyzed how they could be transformed into a more systematic schema representation.
- c) I gathered the available set of industry standards for storing and retrieving biodiversity data and examined for the purpose of developing the actual MorphBank data repository.
- d) Using the analysis from the previous step, I created the initial MorphBank schema that automatically maps data items to their association semantics.
- e) Finally, I devised a new method for the identification, storing, and retrieval of biodiversity annotations.

The validation of the effectiveness of this research is in the implementation and use of the system. MorphBank contains over 100,000 data items and over 60,000 images contributed by over 50 scientists from institutions such as Florida State University, University of Florida, University of Kentucky, Yale, Johns Hopkins, and the

Smithsonian. The MorphBank web site is accessed several hundreds of times each day and usage continues to grow with a steady increase in users, groups, and images.

My particular contribution to this research involved the investigation for improving semantic associations in annotation as well as managing the production of the software. At the start of the project, it was not clear what needed to be accomplished in order to meet the objective of the research. The actual outcome of the research far exceeded all initial expectations. As the research and development proceeded, new ideas were cultivated such object collections and a complete set of complex search/browse functions.

I led a team of programmers and functional analyst in the analysis, design, implementation, and deployment of the MorphBank system. My direct contributions to the project included the following activities:

1. Analysis of the problem

- a. Analysis of the original MorphBank version 1.0.
- b. Analysis of data requirements and gathering of initial MorphBank requirements.
- c. Research of the current state of knowledge of annotations in scientific systems.
- d. Research of available taxonomic name servers.

2. Modeling

- a. Creation of the MorphBank security model.
- b. Creation of the MorphBank data model and schema.
- c. Creation of the semantic association annotation model.

3. Project Manager

- a. Leadership of the design team for the MorphBank system.
- b. Management of the production of MorphBank version 2.2 and 2.5.
- c. Procurement of hardware and software licenses.
- d. Management of the MorphBank NSF/BDI grant under the direction of the Primary Investigators.
- e. Oversight of the functional and design review meetings with users and primary investigators.
- f. Presentations of the project at conferences and workshops.

4. Software Design and Development

- a. Design and implementation of the initial MorphBank Administration Model.
- b. Design and implementation of the initial version of the Taxonomic name selection module.
- c. Design and implementation of the MorphBank Annotation Software.
- d. Design and implementation of the initial version of the MorphBank Collection module.
- e. Design of the external search and exposure feature for the release of MorphBank images in response to MorphOBank external references requirements.
- f. Design of the software test plans.
- g. Contributor to the MorphBank user's manual.

1.1 Problem Definition

Scientist ability to produce large amounts of data has exceeded their capacity to search, processes, and retrieve data effectively. In biodiversity, specimens can be dissected, cataloged, photographed, analyzed, and stored in a variety of media. Much of the detailed knowledge of these specimens is still kept in personal journals, scientific logs, hand-written notes, and human memory. Such informal methods of storing and retrieving information represented a problem when other biologists attempt to search for biodiversity subject matter.

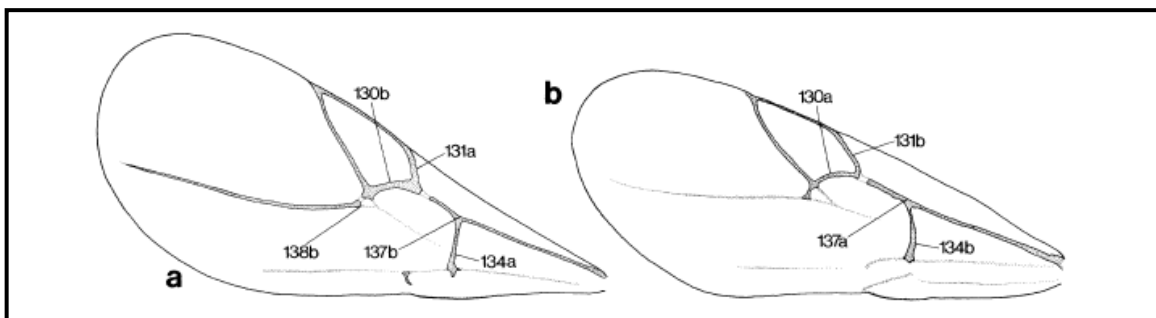


Figure 1-1: Left Fore View, Female a. *Aulacidae phlomica* b. *Synergus crassicornus* [LiRo98]

Figure 1-1 [LiRo98] represents an example of how images are physically annotated and displayed side-by-side for comparison. The figure shows two wings taken from different species of wasp. Image “a.” is a drawing of a wing from a wood boring wasp (*Aulacidae phlomica*) and image “b.” is a drawing of a wing taken from a parasitic wasp (*Synergus crassicornus*). Although they look similar, the characteristics that make the two wings distinct are annotated on the image with labels (130a, 131b... etc). For other biologists to view the comparisons they would have to physically examine the annotated images and attached notes. Scientists have limited methods to discover and view such information.

To solve the problem of searching and viewing such semantically rich information, MorphBank stores images such as that in Figure 1-1 separately and the labels that identifies the characters and states in the form of annotations. With the capability to search images for associated data in MorphBank, scientists can find all related information on an images or any other object.

Attempts at sharing information across the internet have been made using various formats such as XML documents. “However, this new technology has also increased the complexity of these systems and introduced more chaos in the interim” [Shann48]. Most of the middleware applications designed to integrate distributed information systems still assume the underlying schema and functional applications are well designed for use with web services along with an affective user interface. Many biodiversity systems are not designed to export their data to other heterogeneous systems using such middleware applications.

Several efforts have addressed the problem of scientific collaboration. Highly distributed heterogeneous information systems such as data grids [FoKe98], are a response to this problem of information sharing among scientific researchers. In a perfect world; companies, organizations, universities, and interest groups would come together and develop standard methodologies to search and exchange data in an affective manner. There are currently over 20 groups working on distributed database technologies with no centralized agency coordinating their efforts [Oldfi03]. The computer industry would prefer proprietary solutions to the problem of information sharing which would tend to be more portable. Scientists on the other hand would prefer to develop their own individual

solutions that would fit their unique research approach and implement the solutions using open architecture protocols and any available standards.

Sharing of information over these heterogeneous systems becomes difficult and finding information among the different systems is normally not attempted. The project directive from the National Environment Research Council (NERC) Data Grid [Onei02] project illustrates the goal of such a feature: “To provide an easy-to-use interface supporting the discovery, pre-processing and visualization of heterogeneous data from heterogeneous sources for earth scientists.” Although this particular quotation is in reference to the Earth Science data grid, it clearly summarizes the goals of large distributed information systems. The amount of information available through electronic means continues to increase at alarming rate doubling every 5-11 years (depending upon the source referenced). Companies, organizations and educational institutions have a varying array of hardware and software tools at their disposal to create research environments. Therefore, their solutions tend to vary a great deal even within the same disciplines.

Large volumes of information are easy to produce. However, the discovery of information in such large volumes is often difficult and they are often not easy to use or manage. Challenges regarding data grids or any large distributed information system can be categorized in the following areas:

- a) Finding information among the vast universe of data repositories.
- b) Making the information available to interested parties in a secure and reliable fashion.
- c) Aggregating the data into information that can be processed by machines.
- d) Processing the data into information in a reasonable amount of time.
- e) Storing the data in a retrievable media.

1.2 Research Goal

The goal of this research is to develop and implement methods for scientists to search annotated databases for information they need to support their work and collaborate with the scientific community. A single research program is not capable of solving such complex tasks but steps were taken to make a contribution to the area. In an effort to increase the ability to find and store annotated metadata on computational grids the MorphBank Research group developed methods to store, organize, retrieve, and

annotate semantically rich biodiversity data. Separate organizations have different methods and architectures for storing data. However, they're working on similar problems or they have common areas of interest and they wish to share data, information, and resources. Many times, researchers must have detailed knowledge about the information they wish to use such as: (1) what specific data is needed, (2) where the data is located, and (3) how to retrieve the data. Once obtained, the scientist must understand how to transform the data into a useful format. What is required is a single design that can be efficiently implemented on any system that allows heterogeneous systems to locate, transform, and extract desired information for a given.

Ontologies represent a community consensus among the participants and there are usually social pressures against changing that ontology. Problems occur when new participants desire to introduce new definitions and strategies into the system that contradict or alter previous ontologies [Cant04]. Annotations, specimens, and taxonomic descriptions represent an ontology of a group and when someone wishes to use an alternative ontology there is a paradigm shift. MorphBank is designed to handle this shift in ontologies by allowing groups and individuals to develop and use their own.

What is Ontology? One of the clearest definitions I have found for the word "Ontology" is a specification of a specialization [Gruber93]. An Ontology is an item as it relates to the context for which it is used. An example of the problem can be show by searching the World Wide Web for information about a "wing" as it is used in the context of insects. I performed an internet search using the Google™ search engine to find an insect wing. The term "wing" was entered and all of the stored web pages that contained the word "wing" were displayed in order of their discovery. This particular search revealed web pages about music groups, building additions, hockey teams, hockey positions, politics, political groups, radio stations, sports shoes, and a host of other subjects not related to insects. It was not until the 40th web page that an item about a wing of an insect was show. This shows that depending upon the context in which it is used; the word "wing" can take on several definitions.

In the context of distributed heterogeneous systems, ontology refers to the exact context and definition of named data item pairs within a schema. This triplet referenced means the paired objects, their association, and context of the association. In an attempt to address this problem, the biological community has adopted the Darwin Core Standard

[AlRuAn03] set of definitions and should contribute greatly towards the understanding of the individual data items. The standard continues to evolve and future versions of MorphBank will incorporate the new standards into its schema.

This dissertation will focus on the problems of storing informal information systematically and providing a collaborative environment that, until now, was not available. The needs of the biodiversity scientific community for a semantic association annotation model were evaluated and analyzed. A more detailed information model has been developed into a formal schema incorporating any industry and community standards. An annotation model was developed that will allow scientists to make complex associations using name-value pairs. The actual system has been designed, written, and tested using actual biodiversity scientific data.

1.3 Annotation of Metadata Relationships

This research identified the primary relationships among the different objects and their content called “semantic associations”. In most scientific disciplines, the gathering and analysis of data represents a significant portion of the activities involved with the research. The data is then stored, analyzed, transformed (if needed), cataloged, and annotated in a variety of formats ranging from fixed length flat files to relational databases. Data objects have associations that must be identified through semantic descriptions and annotated in order to provide validity to the scientific research community. Informal data repositories tend to be unorganized into hand written ledgers or log books and contain no patterns needed for quick and accurate searches. Recording, understanding, and retrieving information contained in most scientific annotation is one of the major challenges facing the science today.

The following vignette from Ian Foster [Foster02] illustrates the importance of understanding these relationships: “I want to apply an astronomical analysis program to millions of objects. If the program has already been run and the results stored, it will save weeks of computation”. This research focused on observing the relationships among the different metadata items, organizing the data in a logical fashion, and annotating the information to minimize the amount of searching that must be done to discover the correct data. There are current efforts underway to develop similar capabilities. For example, the Chimera project [FVWZ02] has coupled with other data grid services to

enable the creation of new data by executing computation schedules obtained from database queries and managing the distribution of the resulting data. This research project will extend this by annotating the actual transformation instance itself and providing a mechanism to search for and discover the information concerning data transformations.

In order to make this research successful, the information contained in an annotated database must be accessible through the World Wide Web to all participants within a community using standard and open architecture methods. Most organizations will store information in a proprietary manner using various formats such as large flat index sequential files, XML databases, relational databases, or proprietary database systems. Therefore, it would be unreasonable to expect all participants to maintain a duplicate environment capable of reading and storing another organization's data. MorphBank is a web base information system with a complete set of tools that can be accessed using only a web browser. It is currently being used for the storing and retrieval of biological image collections from scientists around the world.

1.4 Verification of Research Objectives

The research will be verified if a semantic associative annotation tool can be developed and successfully used in conjunction with a biodiversity informatics system by the scientific community. It is important to consider that the primary objective of this research is to reduce the time scientists spend in collaboration of their findings and to improve the accuracy of information searches on large scientific systems. An implementation of the new annotation method and database query methods must be developed and deployed to a group of research professionals to ensure the tools are functionality viable. The Annotation trials formed part of the basis for the conclusions of this research.

1.5 Research Accomplishments

The accomplishments of this research went beyond the initial expectations. The requirements of the biodiversity community were documented early in 2004 and identified data items, relationships, definitions, standards, and the current state of development in biodiversity systems. Using this information, a practical data model was developed that was both well defined and flexible. Several individuals and organizations contributed to the development of software prototypes that would form the basis of the system design. Current annotation technologies were reviewed and incorporated into the

MorphBank design. The design was produced into a commercial quality site capable of being used by biologists for the purpose of collaboration of their research. MorphBank version 2.5 has developed into a system that makes biologists more affective because it adds value to digital images by associating semantic information into a single collaborative environment. Feedback from the first year status report to the National Science Foundation supports the accomplishments of this research.

1.6 Research Activities

This research started with the formation of the MorphBank research team early in the year 2004. At this time, the original MorphBank system was analyzed and the initial system requirements were identified. Conference and journal articles were gathered on functional requirements for biodiversity and phylogenetic databases. Additional requirements were added through informal interviews of scientists involved with various significant interest groups such as the HymATOL (Hymenoptera A-Tree-of-Life) group and the TDWG (Taxonomic Data Working Group). The database schema was designed and published for peer review. The initial design of the system was developed and presented at several conferences and working groups for feedback. MorphBank version 2.2 was released in March 2006 and presented at the ATOL (Assembling the Tree Of Life) at Duke University. Version 2.5 was released in July of 2006 in time for the FSU Herbarium Annotation trials and included an improved semantic association system as well as an object collection capability. MorphBank version 2.5 also has a prototype of an XML upload and search capability to test the feasibility of developing an extensible schema system.

1.7 Research Challenges

In communicating to scientists in the meteorology research community, particle physics, and satellite imaging, it was discovered that within each community there is a considerable amount of agreement on standard formats and methods as well as transformation models. However, external the discipline there appears to be a problem. Discovering information in another discipline where one may not have specific knowledge of data formats, file locations, or ontology is a problem. The problem was addressed by adopting the Darwin Core Initiative data standards as a naming convention for the MorphBank data items. Other standards, such as ABCD (Access to Biological Collections Data) were also examined and partially adopted. The MorphBank research

team also adopted the Life Science Identifier (LSID) model as a globally unique identifier for exposing Resource Descriptive Framework (RDF) MorphBank metadata XML documents to external collaborators.

In the case of insect Biology, there exists a fair amount of disagreement on the classification of species and genus among the professional ranks. The issue is whether these differences can be stored logically in a relational manner and annotated to facilitate complex but accurate search techniques. I researched the different taxonomic name servers that were available and recommended to the MorphBank research to adopt the Integrated Taxonomic Information System (ITIS) as their primary taxonomic name server. This was the most stable and complete taxonomic name service available at that time. To provide flexibility, I created a mechanism to add new names to the local copy of the ITIS database and allowed users to annotate specimen data for determination of a specimen.

Annotation of transformation data can be done manually in an informal manner. However, given the volume of data, this is not practical. Many functional areas do have methods and procedures for the automatic identification of metadata and transformation information but these methods would not be useful in other disciplines. MorphBank provides the ability to create complex collections as MorphBank objects that form associations among other objects within the information system. These complex associations form a semantically rich annotation environment for the discovery of information.

The major challenge of the research project is the magnitude of the work that must be accomplished to prove the results. In order to gather sufficient information to form a conclusion, a commercial quality system must be developed, documented, and released for public use. Additionally, a trial of the annotation software must be conducted using actual biodiversity data. MorphBank version 2.5 was released July 29th, 2006 in time for use in a remote annotation trial for a herbarium collection organized and conducted by Dr. Austin Mast of the Florida State University Department of Biological Sciences. Dr. Fredrik Ronquist is planning a similar annotation trial of hymenoptera specimens at a later date. MorphBank is populated with over 100,000 data items and over 60,000 images. A biodiversity database of this magnitude, complexity, and type has never been completed before.

1.8 Chapter Summary

Chapter 2 is dedicated to the work accomplished during the literature search, the discovery of opportunities for MorphBank to improve on current activities, and a forum for the explanation of the more specific topic of annotations in large scientific databases and distributed environments. The primary purpose of this chapter will be to expose the reader to the recent development of the benefits of annotations and the problems that have been generated by the increase capability and capacity of today's networks and computer systems.

Chapter 3 will explain the functional requirements of the MorphBank Biodiversity Image Database which is needed to support semantic annotations. Topics will include a review of data requirements, security, integrity, and annotations requirements.

Chapter 4 will be an in-dept explanation of the MorphBank conceptual model and the underlying design features that were developed to support information discovery and annotations. This complex database was designed in part to support informal data in the form of annotations for the sole purpose of importing a variety of data in a well-formed manner to facilitate the fast and accurate discovery of information. It is important that the reader understand the complexity of the relationship between MorphBank and annotations.

Chapter 5 will address the implementation issues associated with semantic annotations in an ad-hoc environment. An approach will be presented that recommends a method for conducting complex queries efficiently on ad-hoc annotations stored in MorphBank.

Chapter 6 will describe the trial remote annotation trials conducted with the herbarium and hymenoptera determination trials with the Biological Sciences Department at Florida State University.

Chapter 7 will be the conclusion of this document. This chapter will address the findings of the research and additional issues that should be explored.

CHAPTER 2

ANNOTATIONS IN LARGE SCIENTIFIC DATABASES

This chapter will present examples of the current state of research in large scientific databases and the use of annotations for adding value to semantic associations. Additionally, this chapter will also identify opportunities where MorphBank can advance the technologies used in annotations. This chapter will provide readers with a sufficient background to understand the purpose of semantic annotations in biodiversity systems.

There are problems with current state of research and development in biodiversity information systems and annotations. During the literature search for annotations and biodiversity systems, different definitions of annotations were discovered. The following definition represents a consensus among the different uses:

“Annotation: A combination of comments, notations, references, and citations, either in free format or utilizing a controlled vocabulary, that together describe all the experimental and inferred information about a gene or protein. Annotations can also be applied to the description of other biological systems. Batch, automated annotation of bulk biological sequence is one of the key uses of Bioinformatics tools [NG00]”.

This definition of annotations is very broad in the sense that an “annotation” can literally take on different definitions depending upon the application or discipline that is using it. The role of annotation in scientific and scholarly work has historically been associated with laboratory and experimental notes from trusted sources. Different communities define annotations to mean different things. For example, the biodiversity research group refers to an annotation as markers on DNA sequencing data. Often researchers want to annotate different types of digital data (text, binary data, images, audio, video), but there still exists the question on how best to associate the annotation to the data, along with how to search for the annotations and display them.

2.1 Automatic Semantic Annotation

Any type of data entry can be time consuming, tedious, and prone to errors. Therefore, annotations should be as intuitive and as fault tolerant as possible. Alexiei Dingli wrote a paper on Automatic Semantic Annotation using Unsupervised Information Extraction and Integration (UIEI). The UIEI is an idea for extracting information from a

database or other large repositories (such as the web) and automatically adding annotations to scientific data [DCW03]. Most of the current technology is based on human input and knowledge for annotation and is very often a manual process which can be very expensive [Hsm01]. The annotation trials using the MorphBank system have shown that getting the users to annotate thousands of images can be difficult. Annotation requires the commitment of world renowned scientists to sit down and perform data entry.

Semantic annotation is the process of inserting text or markers into the data that have self defining ontologies. In the case of MorphBank, there is a very complex set of interrelated data items that associate the specimen with their related images, localities, groups of individuals interested in the entity, publications, taxonomic descriptions, and image metadata that does not require manual data entry

In an environment where there are different definitions depending upon the context, annotations can be placed on different objects depending upon their ontologies. One of the major obstacles that faced MorphBank was the establishment of the meaning of data items and the establishment of a common ontology for the database. These definitions are required because they describe the concepts and relationships that occur in their respective disciplines. These definitions represent the language used by the different communities to communicate. Basically it describes the domain in which MorphBank users are working. Users of the system search annotations in much the same way that people use a web browser but in a more sophisticated manner using tools that use the system definitions in the criteria. “Producing methodologies for automatic annotation of pages with no or minimal user intervention becomes therefore important: the initial annotation associated to the document loses its importance because at any time it is possible to automatically (re)annotate the document and to store the annotation in a separate database or ontology. In the future Semantic Web, automatic annotation systems might become as important as indexing systems are nowadays for search engines” [Dcw03].

In later chapters, the idea of object associate will be presented that will allow the use of reiteration in data association. Since MorphBank creates objects of different types by inheriting a common base object, we can create a complex set of relationships as the database is built. Additionally, we can search the database to find a related data items. This is a type of automatic annotation that does not involve data entry. As relationships

are built among the different objects in MorphBank we can search for localities of specimens, related specimens, related annotations, supporting publications, external references, objects in related collections, or associated phylogenetic characters.

The association to the internal objects of the system limits the domain and makes the application feasible. For example a technique called Information Extraction (IE) has been used to reduce the burden in some semantic web annotation tools [Vargas02], [Hsms01], and [Cdpw02] and used to crawl the Web for harvesting domain specific information [LeGl01]. In the case of MorphBank, the amount of automated annotation that can be accomplished is somewhat limited as it is in many scientific fields. The types as well as formats of images and techniques used among the different scientists are not standard and vary a great deal. Even among the same research groups, ontologies can be difficult to establish. If there are many different documents and the annotation to be performed very detailed then the annotation process may require a substantial amount of manual work. Scientist may not undertake the task of annotating images in MorphBank because of the amount of work. Therefore, there is the need to make the annotations as efficient and straight forward as possible.

2.2 Gene Sequence and DNA Annotation

One of the most common definitions of annotation is used by the Gene Sequence and DNA research community. Three representative projects (Apollo Genome Annotation Curation Tool, DNannotator, and the Humane Genome Project) were examined as part of the literature search in this research project. In the biology community, the term *annotation* many times references the addition of markers or comments to gene sequence or DNA strand data. Genome annotations are defined formally by the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do [Moun00]. An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it. This particular annotation technique is very specific to that discipline and has precise definition.

Apollo Genome Annotation Curation Tool: The Apollo Genome Annotation Curation tool was developed as a collaboration between the Berkeley Drosophila Genome Project and the Sanger Institute in Cambridge, UK [Lewi02]. Apollo allows researchers to explore genomic annotations at many levels of detail, and to perform

expert annotations in a graphical environment. The tool was used by the FlyBase biologists to construct the annotations on the finished *Drosophila melanogaster* genome. The Generic Model Organism Database (GMOD) project has adopted Apollo as its annotation workbench. Apollo is an open source Java application that can run on Windows, Mac OS X, or any UNIX compatible system including Linux and Fedora. [Lewi02]. Figure 2-1 shows an example of an annotation of gene sequence data using the Apollo tool. While a powerful and very sophisticated annotation tool, Apollo is very specific to small class of annotations and at the writing of this document has a limited capability to analyze the data. Additionally, the annotations are made directly on the graphical data and are not separated. This presents a problem for MorphBank because the images are usually considered to be specimens and cannot be modified. Therefore, annotations must be separated from the object.

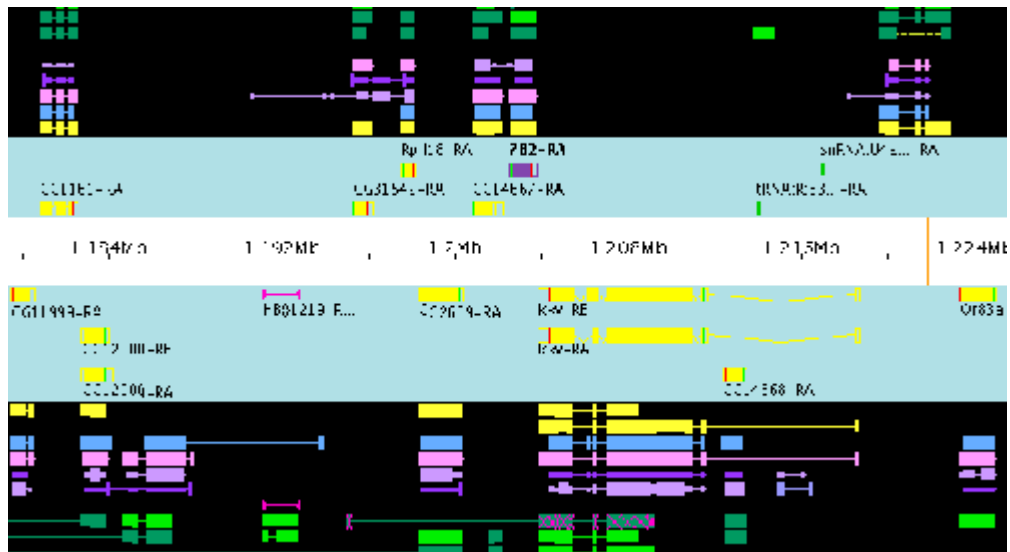


Figure 2-1: Annotation of Gene Sequence Data from the Apollo Genome Annotation Tool. Source: <http://www.fruitfly.org/annot/apollo/>

DNannotator: DNannotator represents a class of systems that combine several tools for both the Annotation and analysis of the data. MorphBank adopted the same philosophy of including the analysis and discovery of data in the same environment for the convenience of the scientists. DNannotator clearly defines annotation as notes on genome sequence data but goes further than Apollo and provides the ability to segregate

the data from the annotation and provides additional tools that allow for analysis and discover that is compatible with the their specific annotations.

Users can customize their own entries by supplying their own annotation source data. This data can be any of the de nova annotations using SNPs, genes, STSs, oligos etc., and their preferred target gDNA sequence for annotations. Although the actual region of annotation is limited to fewer than 30MB for a genomic region, this does not appear to be a severe limitation with the software. DNannotator is a supplement to public annotation efforts such as NCBI's Map Viewer, UCSC's Genome Browser or Sanger's Ensemble. With DNannotator, users can merge sources of public genome annotations and individual findings onto the genomic region of interest [Liu04] [Liu03] [NGM02].

Human Genome Project: The Human Genome Project has been tasked with mapping human DNA. There have been no real standards for annotation in the biology community. Annotation is individualized for each laboratory and scientist. In fact there are numerous annotation toolsets available for download. Another problem discovered early in genome annotations is that genomic features are easy to miss or misinterpret. GenBank entries themselves are annotated very unevenly, depending on the knowledge and interest level of the sequencing lab and once completed it cannot be changed. GenBank is not curated: entries provide only suggestions for genomic features such as promoters, alternative splicing of mRNAs, pseudo-genes, tandem duplications, and homology.

Most annotations need updating to reflect the new information gained through rapid accumulation of genomic sequences. The Human Genome Project performs this task in an unorganized ad-hoc manner making discovery of new annotations difficult. The authentication of neither the author of the annotations nor the reliability of the source is verified in most annotations. Additionally, there is tremendous synergy for improvements in distantly related annotations. However, experience shows that scientists do not have time to update existing annotations and the data is soon out of date. [Liu03] Given the ad-hoc nature of annotations and the diversity of the field, trustworthy expert annotations remains an idea yet to be implemented.

2.3 Semantic Annotations in Image Collections

Since MorphBank is primarily a biodiversity image and phylogenetic database system, the annotation of images is very importance. In this section, we review the work accomplished to date of research that supports the annotation of large image collections. Laura Hollink, Guss Schreiber, Jan Weilemaker, and Bob Wielinga wrote an article on such a topic in 2005 titled *Semantic Annotation of Image Collections*, where they discussed the knowledge based aspect of such a system and the links between the ontologies. [Holli04]

The work referenced is a follow-on effort to previous research on annotation and search of a collection of images for primates [SDWW01]. The subject of the annotation is described through a series of statements that is in-turn associated with an image in the database. The image itself is not altered. A Resource Descriptive Framework (RDF) Schema is used to represent the data. Figure 2-2 shows RDF schema and the overview of the approach of the study.

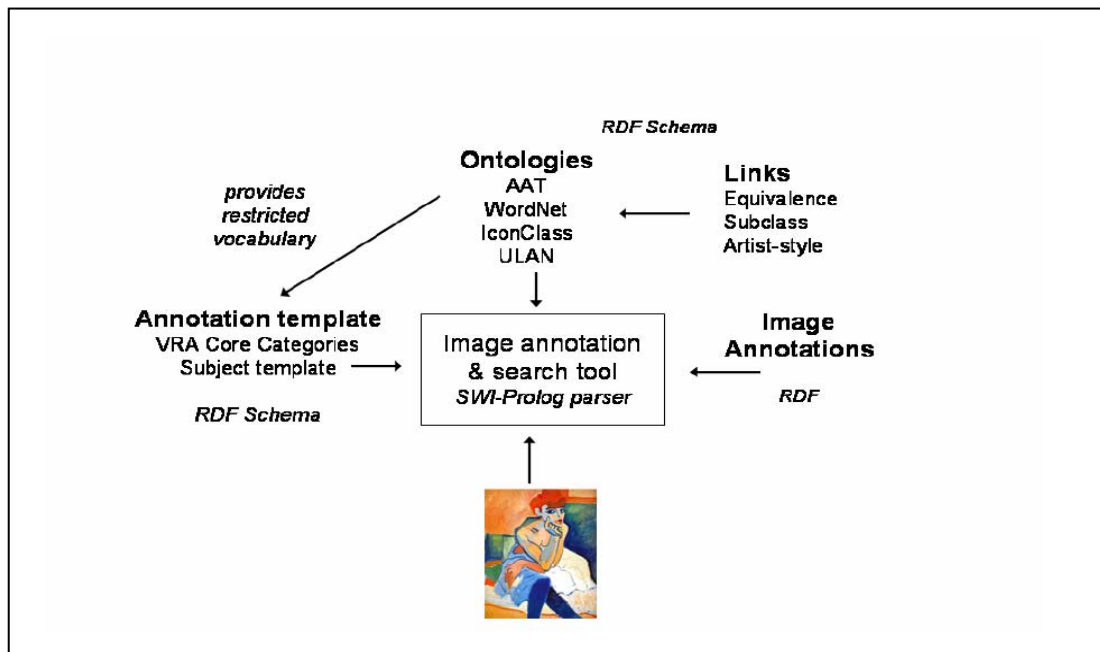


Figure 2-2: RDF Schema over of the tool that generates an annotation template for the displayed image. [Holli03]

Key to the linking of the image, ontologies, and annotations is the development of a thesaurus that allows the Prolog based system to link the different statements together

with an image. Prolog was used because of the ease at which logical expressions can be proven. Because the software is tied to a system that requires such logic, each annotation statement must have a link to at least one *agent* (image) and may have several references to different thesauri maintained by the system. The example given in the paper on annotations using the RDF schema is illustrated by a painting by Chagall titled “The Birthday” (Figure 2-3). In this painting; Chagall kisses his wife and gets flowers from her for his surprise birthday at his studio. The meaning of the annotations on this particular painting can be described using the statements are referenced in Table 2-1.



Figure 2-3: “The Birthday” by Marc Chagall [Holli04]

By using several different ontologies (ULAN, WordNet, AAT, etc) the system can pull together different meanings for the same word or phrase. The user is then able to select the correct meaning that should be applied to the annotation. In the case of the previous example, both ULAN and WordNet have definitions for the phrase “wife” or “wives”. One could consider merging the different ontologies but the complexities of such an operation are prohibitive and would not add any value. Additionally, the users of the system would like the option of reusing different ontologies rather than modifying

them. Despite the use of ontologies and thesaurus, annotations are still opinions and are prone to interpretation and errors. For instance, just by examining the painting by Chagall in figure 2-3 it is not evident that Chagall's wife was giving him flowers. This knowledge was only known by someone familiar with the history of the actual painting.

This approach in annotation of image collection has met with some success at the experimental stages. Small databases have been built and the researchers are able to develop annotations for images (mostly in art collections) where phrases can be stored and searched using a fairly sophisticated method. There still exist the problems of applying this method to a large scale database and a complex set of ontologies. Similar to MorphBank, disagreements in ontologies and approaches in annotations will continue to be a problem. The use of a common set of thesauri to represent the ontologies and an open approach to annotations appear to make the system feasible for use by the art community. The experiments with the RDF schema were generally well received but the limitations in expressivity were noticed. RDF is intended to be a generalized method of exposing data and as such has limitations when being used by a set of individuals who require a great amount of detail at an expert level. A modification of the RDF schema or the inclusion of an additional scheme indigenous to the specific field seems warranted.

**Table 2-1: Thesaurus Reference, Source:
<http://www.cs.vu.nl/~guus/papers/Hollink03b.pdf>**

<u>Heading</u>	<u>Phrase</u>	<u>Thesaurus</u>
Agent:	"Chagall, Marc"	(ULAN)
Action:	"kiss"	(WordNet)
Recipient	"wives"	(AAT)
Agent:	"woman"	(WordNet)
Action	"give"	(WordNet)
Object	"flower"	(WordNet)
Recipient:	"Chagall, Marc"	(ULAN)

Image annotations using only web/browser technology as limitations. One of the early requirements established in MorphBank is the limitation of performing annotations on images and associated data using only the capabilities of a web browser and

JavaScript. The requirement was established for several reasons. The first reason was that MorphBank must be made available to the widest possible audience of scientists and public users. This restriction prevented the development team from producing client based applications to perform more sophisticated image annotations. Such a tool that was investigated by the research team was *ImageJ* developed by Larry Reinking, Department of Biology at Millersville University. This tool was developed using Java on a Microsoft Windows machine that incorporated a wealth of capabilities including annotation, area and location designation, labeling, measuring, and plug-in capability. Like other similar tools, ImageJ has several drawbacks: (1) the annotations actually modified the image, (2) there was no capability to query and extract data from an existing database, (3) there was no capability to communicate with an existing database to deposit the data, and (4) it required that each person using the annotation tool install and maintain a new piece of software.

2.4 Challenges of Annotations on the Web

MorphBank was designed to allow biologists the ability to annotate images and other biological data using web technologies. Annotations themselves can be a very useful mechanism that supports a number of useful functions within a given community. However, the current state of web technologies limit the utility of a web based annotation tool. This section discusses the short falls and how future changes might make a web based annotation tool more useful.

One of the current pitfalls of the web service/client infrastructure is in the support for transparent interception of client-server transactions. [VaPa99] The limitations and inconsistency in today's web browsers are readily apparent. Even with these limitations, users are more likely to accept the problem of customizing their applications on the client side with scripting technologies for the added capability. Server or proxy side applications have more difficulty in scaling capabilities than a client. MorphBank does have some capability to intercept web pages for intermediate computing but these are special cases.

The use of computer technology (client or server side) for digital annotation of electronic images will always be less flexible than manual annotations directly on the printed object. Even more so, HTML or XHTML as a layout tool for digital images is quite poor and standards vary among the different browsers. "For instance, there is no

syntax for sidebar annotations or rendering of annotations on the sidelines or non-use areas on a web page. Similarly, the limitations of graphic manipulations limit annotation of digital images using browser technology. However, proxy-based annotations are usually fairly easy and flexible when compared to client-side methods” [VaPa99]. Built usually with java or JavaScript applications, this method allows for the capture and annotation of objects but still has limitations. Proxy interception of objects does not allow for distinguishing between requests for document or request for subordinate documents.

2.5 Scientific Annotation Middleware (SAM)

The Scientific Annotation Middleware (SAM) system [Myers04] provides significant advances in research documentation and data pedigree tracking that is required for effective management and coordination of the complex, collaborative, cross-disciplinary, compute-intensive research enabled through the Scientific Discovery through Advanced Computing (SciDAC) initiative. The SAM system presents users with a layered set of components and services that provide capabilities for the creation and management of metadata, the definition of semantic relationships between data objects (e.g. pedigree), and the development of electronic research records. [Myers04]

The Scientific Annotation Middleware (SAM) project was designed to address the needs of Grid-based scientific research to federate data and metadata, track pedigree, document research processes, and expose such information to a wide range of services. The key concept behind SAM is “schema-less” data store that can accept arbitrary input and the use of dynamically registered translators to map data and metadata into the formats, schemas, and ontologies expected by applications and underlying data repositories.[MCGS03] Concluded in 2005, the research group succeeded in implementing a the model in a functioning prototype. Agencies that have collaborated with the development of the system and conform to the standards outlined by the project can affectively use the system. To use SAM, participating agents must have developed well organized and fault-tolerant systems at the user level that complies with accepted industry standards.

SAM allows researchers to capture records-related information using an arbitrary combination of tools and to later define how this information should be translated into forms interpretable in other contexts, e.g. into the input format required by a

collaborator's software, the schema of a community database, or that of a records-management tool or automated virtual-data/workflow system. [MyGe03] SAM functions properly if the schemas of the community database and records management tools comply with what SAM is expecting.

Provided the information presented to SAM is in an acceptable form, it becomes possible to view all of the recorded information via a single interface/protocol while simultaneously defining limited views of that data that conform to the conceptual models of particular applications, groups, institutions, or communities. SAM must be able to ascertain the definition of the schema provided by the user and understand the ontology.

SAM is built on the Jakarta Slide content management system and implements the Distributed Authoring and Versioning (webDAV) protocol. WebDAV is an Internet Engineering Task Force standard extension to HTTP that uses XML to encode the content of service requests. SAM allows configurable automated metadata extraction and data translation from binary, ASCII, and XML inputs. XSLT and Binary Format Description (BFD) scripts can be registered with SAM and run dynamically using standard XSLT and BFD engines to extract metadata or create translations and data views. SAM also produces Java Message Service (JMS) events describing all data/metadata access and changes. [MyGe03]

SAM was designed to act as an electronic notebook server compatible with the DOE2000 ELN 5.0 client. Chapters of documents along with individual pages and notes are stored directly via webDAV, and the document associated tree structure is stored as standard properties associated with those resources. In this manner the contents of the notebook are directly available to other webDAV-enabled applications. [Myers04] SAM does not provide the capability for annotation of images at this time. Data from SAM is stored separately from the actual database and is not integrated into the users set of applications.

I investigated the probability of using SAM with MorphBank as an annotation tool. However, MorphBank had not reached the level of maturity required by SAM. SAM is a middleware product that acts as an external interface to the system. MorphBank required the development of an extensive set of lower level components and client software before middleware products could be considered. However, the initial

design of MorphBank included features that would make the use of middleware products such as SAM feasible in the future.

2.6 Chapter Summary

The chapter presents a small sample of the current state of annotation work currently being conducted throughout the community. There are literally hundreds of efforts currently underway developing methods that carry the title of “annotation”. There is a great deal of interest in annotation and as this research has discovered the actual definition of annotation varies greatly from discipline to discipline. GBIF (Global Biodiversity Information Facility) is an example of the need for such a system. GBIF is one of the largest projects underway to document specimens [<http://www.gbif.org>]. Even with a \$3.7 million annual budget and hundreds of thousands of records, GBIF does not incorporate the types of collaborative mechanisms and association semantics inherent in MorphBank.

The effort of this research has been to learn from these efforts and to develop a method of annotation that will allow computational biologists around the world to curate their data in a more efficient manner and to increase distribution of these complex ideas to all interested parties.

CHAPTER 3

REQUIREMENTS FOR A WEB BASED IMAGE AND PHYLOGENETIC DATABASE SYSTEM

We recognized early for the need for an automated repository of biological images for use by research scientists to share their creative works and to collaborate on a variety of fronts. The early version of MorphBank which dates back to 1998 proved the viability of such a concept and established the basic functionality of such a system. Several other systems have subsequently been established since that time. However, the diversity of opinions of research methods has proven to be a major hurdle in collecting a common set of requirements for such a system. MorphBank version 1.0 serves as a baseline for the motivation and essential functionality. This chapter addresses the functional requirements for MorphBank version 2.5+ with particular respect to the focus of this research in the area of object collections, semantic associations, and annotations. We show that by carefully analyzing the actual needs of scientists that we can draw some common areas of interest and yet provide a flexible environment to allow expansion of knowledge and information.

3.1 The Motivation for MorphBank

Two of the major problems facing all biologists today are (1) their inability to share their research with other scientists efficiently and (2) the need to collaborate on projects without extensive travel. Numerous efforts have been undertaken to solve this problem by establishing databases at various locations that attempt to address these problems at least on a limited level. There are literally hundreds of individual biodiversity databases in existence today in various forms of analysis and development. These projects were usually plagued by one or more of the following problems:

1. No data standards are used in the development of the system. Extracting data into other systems was not possible because of the inability to match the ontologies.
2. The development itself was limited in the scope to where the audience for the database was very limited. Development was completed by unskilled software developers using non-standard methods. The data sites were prototypes created to demonstrate a concept.

3. Data in the database was incomplete or inaccurate.
4. Access to the data was difficult (i.e. not web based) or stand-alone.
5. The data displayed was static and not tied to a dynamic database.
6. The database could only be used for a limited range of taxa usually associated with the research of the primary investigator.
7. Limited storage capacity.
8. No security or authentication mechanisms.
9. No automated data analysis or data association capability.
10. Lacked the ability to link with external data sources.

MorphBank is being developed in support of several initiatives around the world including the Assembling the Tree-of-Life (ATOL) project. The general requirements of MorphBank include an interest in locating information about a particular specimen and in identification of species, curation of specimens, collaboration of research with other scientists, research into phylogenetic trees, and other systematic information for a group of organisms. Additionally, MorphBank has the requirement to be used by educators at all levels for assisting in the teaching of subjects in biology and organism diversity.

The overall general requirements for MorphBank include the ability to allow users to search through the site's database for images of biological specimens and to see all related information. This information includes, but not limited to, information on imaging techniques, locality data where the specimen's habitat was located, taxonomic determinations, annotations, phylogenetic characteristics, external reference data, and data on the person(s) who gathered or is responsible for the specimen. There should be a wide range of search and browse features that provide the user a convenient and easy way to find information on MorphBank. Additionally, sharing data with other users should also be convenient. The site's database should serve as both a reference collection of named organisms and a resource for comparative morphological study.

Initially, the system will support a wide range of morphological and phylogenetic study of organisms. The American Heritage Dictionary defines morphology in biology as the study of the size, shape, and structure of organisms in relation to some principle or generalization. Whereas the study of anatomy describes the structure of organisms, morphology explains the shapes and arrangement of parts of organisms in terms of such general principles as evolutionary relations, function, and development. For the most

part, the original MorphBank system version 1.0 included a considerable amount of information on form and structure of insects. MorphBank 1.0 is primarily a digital image repository. Additionally, the general requirements of MorphBank also include a need to store and retrieve phylogenetic data. Phylogenies, i.e., evolutionary histories of groups of organisms. By definition, a phylogeny is a rooted, leaf-labeled tree, whose leaves represent the ancestral taxa. Reconstructing the tree-of-life in terms of the morphological and phylogenetic characteristics of an organism is one of the overall goals of MorphBank.

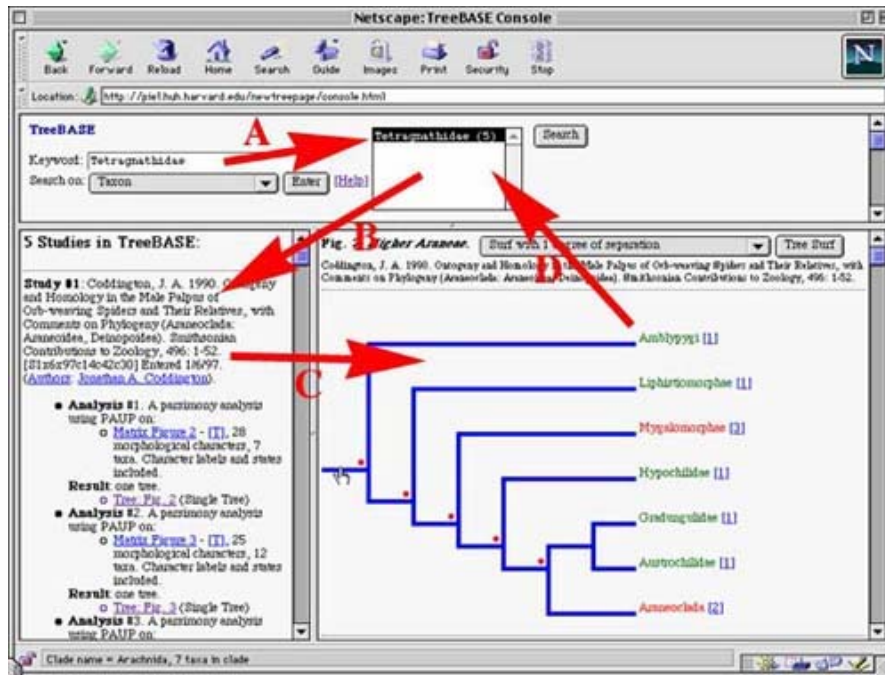


Figure 3-1: TreeBASE Search Console
Source: <http://www.treebase.org>

To fully understand the general requirements of MorphBank, it is necessary that we take a look at some of the previous work in this area. One of the first efforts was TreeBASE started by Michael Donoghue and Michael Sanderson in 1993 [TreeBASE] [Morel96] [Piel96] [PiDoSa02]. TreeBASE stores basically three types of data: (1) published bibliographic information on phylogenetic studies, (2) Corresponding datasets, and (3) the resulting trees in phylogenetic, population, and gene sequences. Besides the number of purposes for TreeBASE, the interesting aspect of the system in relationship to MorphBank requirements are the methods in which the data can be searched.

There are six distinct ways to search for information in TreeBASE: (1) by taxon (i.e. the taxonomic name assigned to an organism), (2) by author or the individual who performed the phylogenetic study, (3) by citation of the publication where the study appeared either by title, author, keywords, or journal, (4) by the study accession number, (5) by the phylogenetic matrix accession number, or by (6) by the structure of the topology of names of the taxa. This basic type of search technique is represented in figure 3-1.

One of the interesting aspects of TreeBase was that it combined the power and stability of a relational database but included a great deal of flexibility by storing the phylogenies as a text field using the Newick format [Olsen04]. This interesting approach was used to allow scientists the flexibility in storing slightly different kinds of data without the restrictions of a tightly controlled relational database. Therefore, queries that concern the structure of the phylogenies are accomplished outside of the relational database.

The rest of this chapter will be organized around the various functional requirements for the MorphBank system. The basic requirements will reference all of the needs of the system starting at version MorphBank version 1.0 through 2.7 which, at the writing of this document, is still under design and development. Additionally, the overall needs of MorphBank version 3.0 will be addressed briefly in the final sections of this chapter.

3.2 MorphBank Security Requirements

This section describes the detailed security requirements for the MorphBank system. Scientists will be using MorphBank for research and collaboration purposes. Although the world can view images and other released data, the ability to add, update, edit, and delete information must be restricted to research scientists and not the general public. Allowing access to data before it has been peer reviewed could result in data being improperly referenced. The relationships of information within MorphBank are highly complex. Individuals who do not have expertise in the area of biology who have access to alter this data could compromise the integrity of the system and diminish the value of the system. This section will describe in detail additional motivation for

MorphBank security and the functional security requirements needed to safeguard the system.

3.2.1 Motivation for Security Requirements

Research biologists require the data deposited in MorphBank have a certain level of security to ensure the results obtained from data analysis is as accurate as can reasonably be expected. During the discussions on security, several practices currently employed by the research biologists at Florida State University and their associates were documented and used to form this portion of the requirements. These discussions were developed into design scenarios listed below:

1. Several scientists use graduate students, undergraduate students, and research assistants in preparation of specimens and data. Access to the database by these individuals should be restricted to a role where span of taxa is very limited and the data is available for review.
2. A level of trust is usually associated with the research developed by scientists. This trust should be reflected in the user's roles and responsibilities within MorphBank.
3. While a dataset is incomplete, it should not be available for review except for a very restricted set of individuals who have complete familiarity with the work.
4. After a dataset is complete, there should be a period of time where the data is available for peer review but not available for world access. This peer review should be limited to a trusted group even if another group has responsibility for the same range of taxa.
5. A user may belong to several different groups and may have a different role for each group. For instance, a person may be a Group Coordinator in one group but only have Guest privileges in another.
6. The Privilege Taxonomic name associated with a person is usually associated with their highest level of competence in an area of biology. This is usually associated with the research group to which person belongs.
7. The Primary Taxonomic name associated with a person is usually associated with their particular area of research. Someone entering data directly related to their Primary Taxonomic name has a higher level of trust than if they

entered data in another area but still within the Privilege Taxonomic name rating.

8. The Secondary Taxonomic Name allows a user to belong and have privileges (similar to that of the Privilege Taxonomic) in another part of the tree-of-life if they have expertise in that area. For instance, a person studying herbivores may have an alternative expertise in botany because of the interest in the vegetation eaten by a particular species.
9. A person should belong to at least one group and this group should be the individual's own group created during account activation. That person would have Group Coordinator status for their own group.
10. An object may be owned by one person and one group at any given time. Only a MorphBank administrator can change ownership for security and integrity purposes.
11. Assume a person can belong to multiple groups. A user must select a group after logging in order to assign ownership to objects created and to verify permissions for access, edit, add, and annotation.
12. MorphBank data is accessed from systems external to MorphBank. Controls should be put into place that prevents unauthorized users from gaining access to objects they are not authorized.

3.2.2 MorphBank Access, User Roles, and Security Requirements

MorphBank is an open web repository of biological images designed to serve the research community and as such the reliability and accuracy of the data is of significant importance. To ensure the data is of the highest accuracy, safeguards must be put into place that will allow scientist to input data into the system accurately and have it peer reviewed before release to the world. Once released to the world, the data is considered of museum quality and must not be tampered with by unauthorized individuals. Expertise in specific areas of biology dictates the privileges a user has within each group to which they belong.

Users of the MorphBank system are divided into to general categories of roles and must be assigned to groups within MorphBank for the purpose of access to restricted data. Anyone can browse or search for released images and data on MorphBank and without the need to login the system. There are other categories of users that have rights

to modify the MorphBank data and as such are required to register with the system with a login name and password that identify them with particular access rights. This section describes those rights and access privileges.

Users registered with MorphBank will have a user account with the system. They will request a unique username from the MorphBank Administration and are given an initial password. They have the authority to change the password and personal information through the personal user maintenance account. All users of MorphBank regardless of version have the same basic privileges in the system to browse and search images on MorphBank. Additional access is granted by membership in groups managed by a Group Coordinator. Groups are identified by a name, the taxonomic range for which they responsibility, and the MorphBank objects for which they own. The taxonomic range is identified by the Taxonomic Serial Number (TSN) as used in the ITIS database system. For instance, a group responsible for images and data in the family Hymenoptera would have a TSN assigned as 152741 and would allow users of that group access to related objects with that TSN range or lower. Users may be members of several groups. If for instance, a user is a member of the Hymenoptera group and also of the *Cephus cinctus* (Wheat Stem Sawfly) they would have access rights to objects owned either group. However, to maintain the integrity of the data, they would have to declare the group they belong to when modifying or adding data in order to maintain data ownership.

Object Ownership: MorphBank will have the ability to track ownership of individual objects within the database. Ownership is comprised of (1) the person in MorphBank that contributes the object, and (2) the group that will have primary responsibility for verification and validation of the data. While the object is under review (prior to the release date) MorphBank users that do not belong to the object's group may not view the data, modify it, nor annotate it. While the object and associated data is being entered and no release date established, only the owner (contributor) of the object may modify it. The Coordinator of the group and Lead Scientists who are members of the group; may view or annotate the object. Once an item has been released, all users (including the world) may view the object and all MorphBank registered users may annotate the data.

Rating of Users: Users are placed in categories that depend upon their area of expertise. To determine the range of responsibilities of a person there are four associated Taxonomic Serial Numbers (TSNs) associated with each user. They are as follows:

1. **Privilege TSN:** This TSN value identifies the highest class of biological entity for which the User has education, training, and experience. When requesting a MorphBank User account, a user will state their highest taxonomic range. For example, a person whose recent research involves wasps, ants, and bees may request their Privilege TSN to be 99208(Class: Insecta).
2. **Alternative Privilege TSN:** This TSN is usually a value in another branch of the Taxonomy Tree that would allow a user to hold scientist or lead scientist privileges in another group. For instance, specialists in the area of plants may have a detailed knowledge of bacteria that affect plants in their specific area.
3. **Primary TSN:** The Primary TSN identifies the user's specific area of expertise. This TSN usually indicates that the user has done extensive research in this area and is considered by his peers as an expert over this range. This TSN range is usually at or below the family level. As an example, an individual with extensive experience in certain parasitic wasps may have a Primary TSN of 150425 (family: Cynipoidea).
4. **Secondary TSN:** Often, users will belong to working groups that have a wider taxonomic range than their own specific expertise. Following the previous example, a person whom does research on Cynipoideas may belong to a Hymenoptera working group and have a Secondary TSN of 152741 (Order: Hymenoptera).

3.2.3 Categories of MorphBank Users

Within each group, members may hold a range of privileges called roles. These roles range from creating new groups and adding users to merely being able to view and annotate unpublished objects. Access rights are controlled by membership into specific system roles organized by groups. A user may have a different role in each group to which they belong. A person who holds Group Coordinator in one group may only be a Lead Scientist in another but at the same time hold the role of Guest Privilege in all other groups. The role that a person has in each group defines their privileges. The initial sets of roles are defined below:

1. **Administrator:** This title refers to individuals who are in charge of the security and maintenance of the MorphBank Database system, hardware, and software. Typically a MorphBank Administrator is one of the developers or agency in charge of the day-to-day operations and security of the system. There are relatively few Administrators. Once a person is assigned Administration privileges, they need not be assigned to any of the other roles.
2. **Group Coordinator:** There is only one Group Coordinator for each group and this person must have held or have been assigned the title of Lead Scientist. As a Group Coordinator, this individual can assign roles and responsibilities of Lead Scientist or Scientists to other members of the group. In addition, a Group Coordinator can also add or remove members from groups. The Group Coordinator can request to the MorphBank Administration that a new group with Taxon Range within the current group be created and recommend a Group Coordinator be assigned. It is important to note that a Group Coordinator must have a Privilege TSN of at least as high (or higher) as that assigned to the group. The Group Coordinator may appoint another Lead Scientist in the group as the Group Coordinator. A Group Coordinator has all privileges of a Lead Scientists.
3. **Lead Scientist:** A Lead Scientist has the ability to add and update views, location information, phylogenetic characters and state data, publication data, add new temporary taxonomic names, and other related support data into MorphBank. A Lead Scientist will have the responsibility for the integrity and accuracy of the data contained with MorphBank and other related data stores. It is important to note that a Lead Scientist must have a Privilege TSN of at least as high (or higher) as that assigned to the group. A Lead Scientist has all the privileges of a Scientist.
4. **Scientist:** This is the basic membership of MorphBank and is anticipated that most users in MorphBank will have Scientist as their highest role. In this role, users will be allowed to add and modify the database prior to release of image and specimen data. They will also be allowed to add image and taxonomic identification annotations to data they have contributed or data that has a taxon range within a group for which they belong. It is important to note that users may

- not add, alter, or annotate data for a taxon range for a group in which they do not belong or on objects for which they do not own.
5. **Guest:** Users with role privileges has access rights to view and annotate data owned by groups for which they belong. Guests may not add or alter MorphBank data other than their own annotations. All users with login access in MorphBank have at least Guest access rights.
 6. **World:** No login name is required for World access. Users are only allowed to browse and search for data but may not make any modifications or annotations.

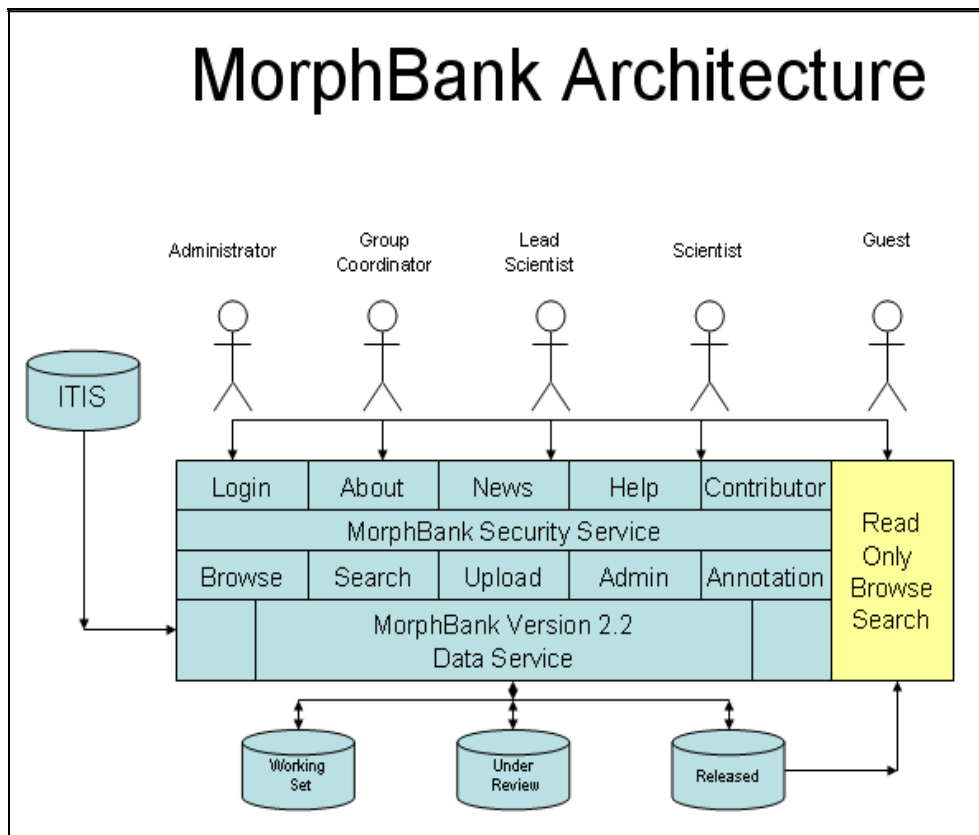


Figure 3-2: MorphBank Architecture

3.3 Data Access Requirements

The main purpose of MorphBank is to provide research biologists with an open access, reliable, and accurate data repository of phylogenetic data 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 [<http://morphbank.net>]. To obtain this, the data in MorphBank must be as

accurate as possible. This section describes the data access and security model that will be used in the software to ensure the data in MorphBank is of the highest quality. There are three distinct categories of data (See Figure 3-2) contained in MorphBank. (1) Working Set, (2) Under Review, (3) Released. They are described below:

1. **Saved:** Scientists and above are responsible for entering data into the MorphBank system. While the data is being entered no release date will be assigned to the individual objects associated with the data. To allow users access with membership in groups, the contributor need only assign a release date.
2. **Submitted:** Once a dataset has been entered and complete, the contributor will submit the item and assign a release date to the objects. Members in that group with roles of at least Guest may now browse, search and review the objects and annotate the images for general use or taxonomic identification. Under specific conditions, the contributor can make corrections to the data.
3. **Released:** Once the data has reached the release date, all users and roles have read access to the data. Users of all groups may make annotations but modifications to the data (including updates and deletes) are not allowed at this time. Group ownership is no longer validated when the data is released. Once MorphBank data is released is becomes a permanent record in the database.

3.4 MorphBank Data Requirements

In this section, a simple data model is proposed that will support the overall functional requirements of the MorphBank system. The schema is represented in a normalized relational database by representing each data object and their main supporting data items in general terms. Although the actual MorphBank database schema contains over 48 tables and several hundred attributes, only the major MorphBank objects that are visible externally are defined in this schema. Figure 3-3 shows the MorphBank basic entity model which illustrates the relationships among the different MorphBank objects.

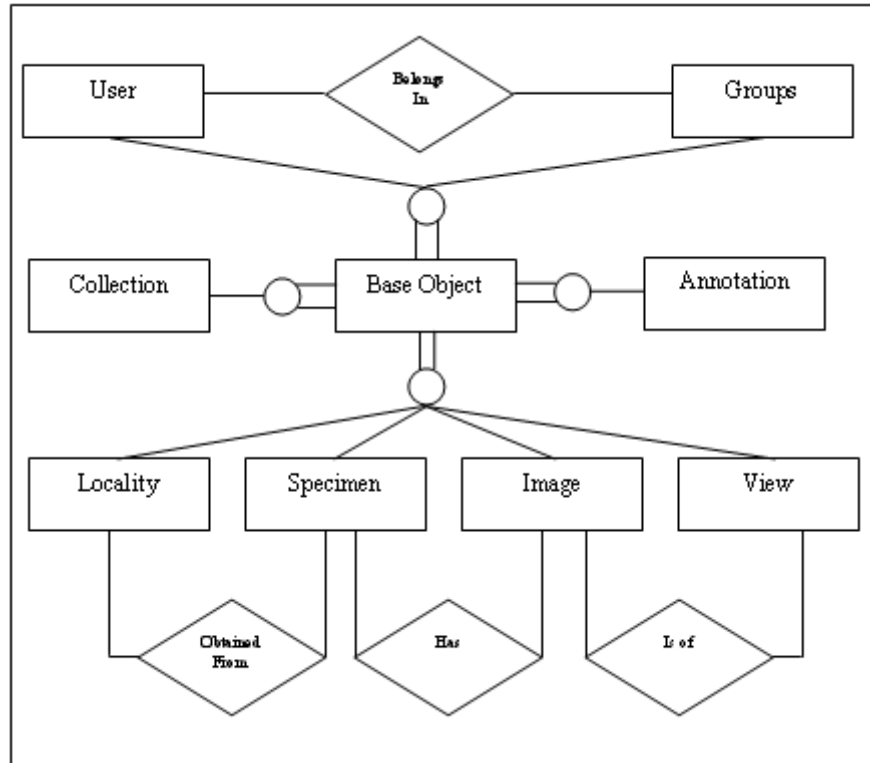


Figure 3-3: MorphBank Object Model

3.4.1 The Specimen-scheme

Specimen = (id, ownership, collectionInformation, locality, determination) The Specimen object represents the central object for which all other objects have either a direct or indirect relationship to. As in all objects, a unique internal identifier is associated with each Specimen. This schema identifies all data that is unique to the individual specimen. For the purpose of MorphBank, a Specimen is identified as a biological entity of interest. Provisions must be made in the system that allow for portions of a Specimen to also be considered a Specimen and this relationship should be represented in the database.

3.4.2 The Image-scheme

Image = (id, imageFile, ImageSpecification, specimenId, instituteInformation, instituteCollectionInformation, taxonomicNames) The original MorphBank website used primarily images as the central source of

information. Similarly, MorphBank versions 2.2 and higher also rely heavily on the use of images. As a morphological website, images and the appearance of biological entities are of the most importance. Views, phylogenetic characters, and image annotations are all directly related to images.

3.4.3 The User-scheme

User = (id, name, affiliation, privileges, groups) The user schema represents those individuals that have access to sensitive or restricted data within MorphBank. Any individual users may have only one account on MorphBank but, because of the advent of groups and roles, they may have multiple roles and privileges in any group.

3.4.4 The Group-scheme

Group = (id, groupName, taxonomicRange, groupCoordinator) The group schema was required to allow individual scientists the right to some privacy while organizing and inputting the data for the collections. Objects within MorphBank have an owner (user who owns and created them) and a group to which they belong. The user must be a member of the group in order to declare an object of that group.

3.4.5 The View-scheme

View = (id, viewName, imagingTechniques, relatedTaxonomicInformation, specimenPart) When entering image data, the user has the ability to declare that the image adhered to exact specification and standards. These standards and specifications are declared in the View object.

3.4.6 The Locality-scheme

Locality = (id, latitude, longitude, minimumElevation, maximumElevation, minimumDepth, maximumDepth, continentOcean, country, localityInformation) Locality information can be sensitive. The exact location of certain endangered species should not be entered into this version of MorphBank. Locality information references to location of where an individual specimen or group of specimens were gathered or observed. The data of the observation or locality is also important.

3.4.7 The Annotation-scheme

Annotation=(id, typeOfAnnotation, objectId, comments, relatedObjects) An annotation is an additional comment or piece of data associated with an object. The design of MorphBank allows for the annotation of any object in MorphBank but as of version 2.5 only images and specimens are annotated.

3.4.8 The Publication-scheme

Publication=(id, publicationType, authorInformation, title, chapter, edition, editor, institution, journal, month, volume, publisher, pages, series) Authority is a very important component of MorphBank. Scientists require the ability to associate with a specimen, phylogenetic character state, or collection with a publication for reference.

3.4.9 The PhylogeneticCharacters-scheme

PhylogeneticCharacters = (id, characterNumber, description, relatedTaxonomicName, sex, relatedCharacter, relationshipType, publicationId) Phylogenetic characters and character/states reference the evolutionary development of organisms. By treating phylogenetic relationships rather than organism traits as necessary and sufficient properties, it is believed that phylogenetic definitions remove conflicts between the definitions of taxon names and evolutionary concepts of taxa. The general method of definition represented by phylogenetic definitions of clade names can be applied to the names of other kinds of composite wholes, including populations and biological species. A character represents a specific characteristic of an individual organism. An example might be the shape of a female wing that belongs to the family of gall wasps.

3.4.10 The PhylogeneticState-scheme

PhylogeneticState = (phylogeneticCharacterId, stateId, description, definingImageId, extremeImageId-1, extremeDescription-1, extremeImageId-2, extremeDescription-2) Directly related to the PhylogeneticCharacters, the PhylogeneticState represents the number of

different variations that can occur in that particular state for that particular taxa. Therefore, a PhylogeneticCharacter may have several states. Examples: (a. slight bend forward, b. prominent vein angled at 36% rearward).

3.4.11 The Collections-scheme

Collections = (id, publicationId, name) Collections are a unique concept in MorphBank. While performing a query, registered users have the ability to identify images they wish to include in a “Collection”. Collections are a group of related objects that have ownership at the group and user level like other MorphBank objects. A Collection may have one or more object associated with it.

3.4.12 The CollectionObject-scheme

CollectionObject = (collectionId, objectId, objectType, title, order)

CollectionObject schema is the table that contains the references to the objects related to each Collection. Each Collection object references the Collection it belongs to by a collection Id and an object identification, type, and order.

3.5 External Object Exposure

Early in the project, the research group decided to design the new version of MorphBank in such a manner that would allow data within MorphBank to be made available using Life Science Identifiers (LSIDs) as a means to expose the objects and RDF (Resource Descriptive Framework) as means to show the information. If for instance, a piece of information about a biological data were sent via an email or other document, the user could click on the link as a URL and your computer would understand that this was an LSID. This would initiate an application to start to resolve this LSID into one or more references for the data to include any specialized applications that were needed. Once exposed, the user would then have access to an RDF document that would provide them with all of the available metadata about the objects. Figure 3.4 shows examples of a complete LSID.

urn:lsid:pdb.org:1AFT:1

This is the first version of the 1AFT protein in the Protein Data Bank.

urn:lsid:ncbi.nlm.nih.gov:pubmed:12571434

References a PubMed article

urn:lsid:ncbi.nlm.nih.gov:GenBank:T48601:2

Refers to the second version of an entry in GenBank

Figure 3-4: LSID Examples, Source <http://lsid.sourceforge.net/>

LSIDs have five parts: the Network Identifier (NID); the root DNS name of the issuing authority; the namespace chosen by the issuing authority; the object id unique in that namespace; and finally an optional revision id for storing versioning information. Each part is separated by a colon to make LSIDs easy to parse. The main requirement for a piece of data to be considered an LSID is that it must be unique and persistent. In order to satisfy the requirement to be LSID compliant, MorphBank required that each individual object within the database be uniquely identifiable. To extend the requirements, the following requirements were identified early in the development of the MorphBank project:

1. Each major object within MorphBank (Specimen, Image, User, Group, View, Locality, Publication, Collection, and that can be exposed with LSIDs must have a unique serial number assigned to it.
2. Each major object will be registered as a baseObject Schema.
3. Each baseObject will have an owner, group, data created, date to be published, and data last modified.
4. Once a baseObject item is created, it cannot be deleted unless all related objects that reference it are deleted or modified to remove the reference.
5. Once an object is made public, it cannot be changed.
6. The requirements to catalog each object and to enforce foreign key constraints within the system. The requirement will allow complex relationships to evolve among the major objects of MorphBank.

MorphBank also has the requirement to expose objects to the world through LSIDs using the RDF (Resource Descriptive Framework) format. The RDF format

integrates several different applications using the XML standard as a method of interchange. The RDF specification provides a lightweight ontology to support the exchange of information on the web [SCHEMA]. Using LSIDs as a means of exposing objects in MorphBank and RDF as an open format for sharing information, other research organizations can gain access to data in a non-restrictive format. Figure 3-5 shows an example of information concerning an image in MorphBank exposed through the RDF format. By using RDF, MorphBank can encode, exchange, and reuse structured metadata seamlessly. RDF allows for metadata interoperability through common conventions of semantics, syntax, and structure. RDF does not include semantics for each community, but allows these communities to define their own metadata elements as needed. RDF uses XML as a common syntax for the format of the documents. By exploiting the features of XML, RDF imposes structure that provides for the unambiguous expression of semantics and, as such, enables consistent encoding, exchange, and machine-processing of standardized metadata. [SCHEMA]. This very powerful capability will make MorphBank a very flexible data repository that will greatly assist in collaborations of the world's biologists.

```

<rdf:Description rdf:about="urn:lsid:morphbank.scs.fsu.edu:morphbank:66007">
  <mbank:specimen
rdf:resource="urn:lsid:morphbank.scs.fsu.edu:morphbank:64282"/>
  <mbank:view rdf:resource="urn:lsid:morphbank.scs.fsu.edu:morphbank:63977"/>
  <rdf:type
rdf:resource="http://morphbank4.scs.fsu.edu:8080/rdf/morphbank#Image"/>
  <mbank:description>Width and Height set</mbank:description>
  <mbank:imageWidth>829</mbank:imageWidth>
</rdf:Description>
<rdf:Description rdf:about="urn:lsid:morphbank.scs.fsu.edu:morphbank:64282">
  <darwin:kingdom>Animalia</darwin:kingdom>
  <mbank:images
rdf:resource="urn:lsid:morphbank.scs.fsu.edu:morphbank:66007"/>
  <rdf:type rdf:resource="http://digir2.ecoforge.net/rdf-
schema/darwin/2005/2.0#DarwinCoreSpecimen"/>

```

Figure 3-5: Sample RDF for an Image[Ricca06]

3.6 Rudimentary Query Requirements

In this section we explain the basic query requirements that will be necessary to support the information retrieval needs of biologists regardless of specialty. Table 3-1 shows the relation name (table) and the symbol that will be used in the relational algebra expression that corresponds to the query. The relational calculus query requirements that follow are based upon the needs of the research scientists interviewed in the initial MorphBank requirements analysis with additional information obtained by study of the initial MorphBank version.

Table 3-1: Relation Symbol

<u>Relation Name</u>	<u>Relation Symbol</u>
Image	I
Specimen	S
View	V
Locality	L
User	U
Group	G
Taxonomic Units	T
Collection	C
Annotation	A

3.6.1 Image Queries

Image query requirements will always return images as a result. Figure 3-6 shows the basic query statements that represent the minimum number of requests a user is likely to make. IQ1, IQ2, and IQ5 are all simple queries that return images based upon simple criteria such as image attribute, specimen id, or view id. The other queries require some type of join operation to extract related information such as user, group, collection, or annotation information. Selecting all images for a specific Specimen (IQ2), selecting images that belong to a collection (IQ6), and selecting images of a specific annotation (IQ7) are of the most importance in version 2.5. In my analysis of the requirements, I discovered that biologists were most interested in selecting images of

specific genus and species which were deposited by certain scientists. Therefore by combining the results of IQ3 and SQ4 (Figure 3-6) we can obtain these types of results.

<p> IQ1: Result $\leftarrow (\sigma_{id=R}(I))$ IQ2: Result $\leftarrow (\sigma_{specimenId=R}(I))$ IQ3: Result $\leftarrow I \bowtie_{id=userId} (\sigma_{id=R} B)$ IQ4: Result $\leftarrow I \bowtie_{id=groupId} (\sigma_{id=R} B)$ IQ5: Result $\leftarrow \sigma_{viewId=R} I$ IQ6: Result $\leftarrow I \bowtie_{id=objectId} (\sigma_{id=R} C)$ IQ7: Result $\leftarrow I \bowtie_{id=objectId} (\sigma_{id=R} A)$ </p>

Figure 3-6: Image Queries

3.6.2 Specimen Queries

The results of a specimen queries not only returns information concerning the specimen but also the primary image. Unlike many of the other relations with the MorphBank schema, there exist some information with the Specimen relations that must be joined with other tables in order to present a view that is understood. For instance, SQ4 of Figure 3-7 shows a query of a specimen for a specific Id that is joined with the Taxonomic Unit table to retrieve the genus and species. Alone, the Taxonomic Serial Number (tsnId) has no meaning to the casual user. Likewise SQ2 and SQ3 join with the User and Group relations to retrieve user and ownership information.

<p> SQ1: Result $\leftarrow (\sigma_{id=R}(S))$ SQ2: Result $\leftarrow S \bowtie_{id=id} (\sigma_{userId=R} B)$ SQ3: Result $\leftarrow S \bowtie_{id=id} (\sigma_{groupId=R} B)$ SQ4: Result $\leftarrow \sigma_{tsnId=R} T$ SQ5: Result $\leftarrow S \bowtie_{id=objectId} (\sigma_{id=R} C)$ SQ6: Result $\leftarrow S \bowtie_{id=objectId} (\sigma_{id=R} A)$ </p>

Figure 3-7: Specimen Queries

3.6.3 View Queries

View queries are of particular interest to biologists because in many cases the views are not only indigenous to a genus or species but may also relate only to a particular body part in a certain angle and of a particular sex. If this knowledge about views are known to the user they can quickly narrow the search for particular images very quickly. Views always return images of specimens. Figure 3-8 shows the applicable view queries.

$$\begin{aligned} \text{VQ1:Result} &\leftarrow (\sigma_{\text{id=R}} V) \\ \text{VQ2:Result} &\leftarrow V \bowtie_{\text{id=id}} (\sigma_{\text{userId=R}} B) \\ \text{VQ3:Result} &\leftarrow V \bowtie_{\text{id=id}} (\sigma_{\text{groupId=R}} B) \\ \text{VQ4:Result} &\leftarrow (\sigma_{\text{tsnId=R}} V) \\ \text{VQ5:Result} &\leftarrow V \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} C) \\ \text{VQ6:Result} &\leftarrow S \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} A) \end{aligned}$$

Figure 3-8: View Queries

3.6.4 Locality Queries

The original query requirements on Locality were to find specimens gathered at the same sight or area. However, Locality took on a slightly different definition. Expeditions for specimen gathering that occurred at the same location but in different years were entered as two separate localities. Therefore, queries for specimens and images given a specific Locality will yield results for a specific expedition. Searches can also be conducted for ranges of latitude and longitude and country or sear. Figure 3-9 shows the basic set of locality queries.

$$\begin{aligned} \text{LQ1: Result} &\leftarrow (\sigma_{\text{id=R}} L) \\ \text{Q2: Result} &\leftarrow L \bowtie_{\text{id=userId}} (\sigma_{\text{Id=R}} B) \\ \text{LQ3: Result} &\leftarrow L \bowtie_{\text{id=groupId}} (\sigma_{\text{Id=R}} B) \\ \text{LQ4: Result} &\leftarrow L \bowtie_{\text{id=objectId}} (\sigma_{\text{Id=R}} C) \\ \text{LQ5: Result} &\leftarrow L \bowtie_{\text{id=objectId}} (\sigma_{\text{userId=R}} C) \end{aligned}$$

Figure 3-9: Locality Queries

3.6.5 User and Group Queries

Most of the queries for users associated with specimens, images, views, localities, groups, and annotations are incorporated with those queries. Since MorphBank is a secure data repository, specific information that only concerns the attributes of a user will only be available to other users. As per figure 3-10, other users will only be able to look up other users and find to what groups they belong. Queries on the Group relation will return a result of information about that particular group including the name, taxonomic range, and the group manager.

$$\begin{array}{l} \text{UQ1:Result} \leftarrow (\sigma_{\text{id=R}} \text{U}) \\ \text{UQ2:Result} \leftarrow \text{U} \bowtie_{\text{id=groupId}} (\sigma_{\text{Id=R}} \text{B}) \\ \text{GQ2:Result} \leftarrow (\sigma_{\text{Id=R}} \text{G}) \end{array}$$

Figure 3-10: User and Group Queries

3.6.8 Collection Queries

Collection queries and Annotation queries form a close relationship in that collections are a form of annotations. Obtaining the results of a query for a collection owned by a group or user will return that collection and all associated object information. Although in MorphBank version 2.5, a collection can only exist of images, the design of the schema will allow for any MorphBank object to be included in a collection. Figure 3-11 shows the collection queries. A user will be able to create a collection from an initial set of objects that are the result of some query. Additionally, a user will be able to add objects to a collection unrestricted by whether the objects are images, specimens, views, localities, users, groups, annotations, or even other collections. A collection will have an order of the objects that can be altered by the owner prior to publication. The owner will also have the ability to give a distinct name to each object and perform sorting and annotation functions on the collection as a whole or any proper subset.

<p>CQ1: Result $\leftarrow (\sigma_{\text{id=R}} C)$</p> <p>CQ2: Result $\leftarrow C \bowtie_{\text{id=userId}} (\sigma_{\text{id=R}} B)$</p> <p>CQ3: Result $\leftarrow C \bowtie_{\text{id=groupId}} (\sigma_{\text{id=R}} B)$</p>

Figure 3-11: Collection Queries

3.6.9 Annotation Queries

Annotation queries are some of the most complex in the system because they incorporate aspects of many of the MorphBank objects. Queries on annotations can return results based upon objects related to specific images (AQ2 Figure 3-12) which are also cross referenced with specimens (AQ3 Figure 3-12). Annotations written by specific users under the ownership of certain groups are also significant.

<p>AQ1: Result $\leftarrow (\sigma_{\text{id=R}} A)$</p> <p>AQ2: Result $\leftarrow A \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} I)$</p> <p>AQ3: Result $\leftarrow A \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} S)$</p> <p>AQ4: Result $\leftarrow A \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} U)$</p> <p>AQ5: Result $\leftarrow A \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} G)$</p> <p>AQ6: Result $\leftarrow A \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} V)$</p> <p>AQ7: Result $\leftarrow A \bowtie_{\text{id=objectId}} (\sigma_{\text{id=R}} C)$</p>

Figure 3-12: Annotation Queries

3.7 Chapter Summary

The topics presented in this chapter represent a comprehensive set of requirements for an image based phylogenetic information system. All aspects of the project to include both the detailed data requirements, security, accessibility, and integration issues were addressed. One of the most important features of the MorphBank system is the ability to seamlessly integrate the different features of the system in a manner that is completely transparent to the user. One of the overall guiding principles behind the development of these requirements was the idea that MorphBank was to be used as an information repository and discovery tool to be used by research biologist.

Additionally, the research team wanted to reduce the amount of time needed for a biologist to use and integrate the system into their day-to-day activities.

CHAPTER 4

MORPHBANK CONCEPTUAL MODEL

The three overall guiding principles behind MorphBank are the reliability of the information contained within the database, the security of the data during the curation process, and the identification and discovery of the complex relationships of the different data objects. These three principles guided the development of the MorphBank data model. In chapter three, the overall general requirements for MorphBank were discussed. They include the ability to allow users to search through the site’s database for images of biological specimens and to see all related information. This information includes, but is not limited to, information on imaging techniques, locality data where the specimen’s habitat was located, taxonomic determinations, annotations, phylogenetic characteristics, external reference data, and data on the person(s) who gathered or is responsible for the specimen. Additionally, the functional need to safeguard the information was also presented. There should also be a wide range of search and browse features that provide the discovery of all data and interrelated objects.

The remainder of this chapter will be used to discuss the individual conceptual models used to develop the MorphBank system and the concepts that were developed to support Semantic Annotations and the discovery of ad-hoc data.

Table 4-1: MorphBank Basic Objects

Object Name	Symbol	Description
Specimen	S	Holds unique Specimen data
Image	I	Holds location of image and related data
View	V	Holds data of the view that belongs to the Image
Locality	L	Location information where the Specimen was observed, raised or collected
Collection	C	The primary keys of related objects
Annotation	A	Additional notes, comments, or data about any of the objects within MorphBank
Publication	P	Reference material used to verify data in any of the MorphBank objects.

4.1 Base Object Relationship

Table 4.1 shows the basic objects that make up the MorphBank information system. Although conceptually these are the primary data points within the system, there are considerably more relations that are required to accurately represent the data. Each object in MorphBank must have an owner who either is registered with MorphBank or was at one time. Additionally, the user that owns the object must declare under which group the object is contained within.

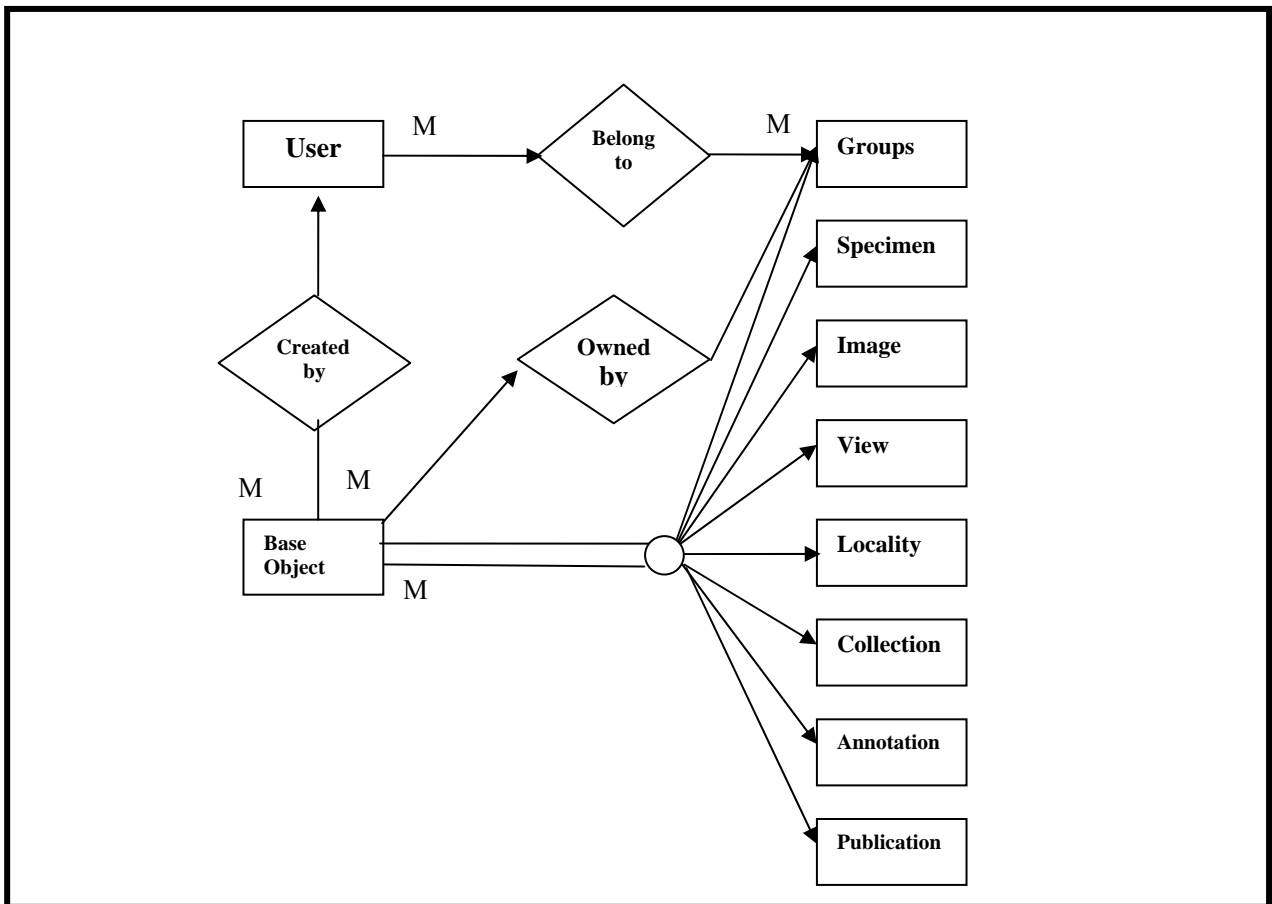


Figure 4-1: User and Object Relationships

The MorphBank model follows a basic object-oriented theme in the division of data items and in defining the relationships of the objects to each other. Figure 4-1 depicts the relationship of the objects in association with the owner of the object (User)

and the catalog that contains the basic information about each object (Base Object) which is inherited by each, and the individual objects themselves. The Base Object of MorphBank denoted by B contains the base set of information inherited by each of the different objects (Specimen S , Image I , View V , Locality L , Collection C , Annotation A , and Publication P). To be valid, each object must represent a proper subset of the Base Object.

Each of the objects in MorphBank are individual sets where any set T with any other set R written as $T \cap R$ is the empty set. This same feature requires that full set of objects (S, I, V, L, C, A, P) is a proper subset of the Base Object B . Let $B(x)$ be the set of set (S, I, V, L, C, A, P) of the Base Objects of MorphBank. Since each of the elements in $P(x)$ represents a proper subset of the Base Object B , then we can state the following:

$$B(x) = \forall_x \exists_y (z \in y \iff z \in x)$$

The Base Object is then a Power Set because all of the Base Objects represent the complete set of all subsets that are allowed in the MorphBank information System and is shown in Figure 4-1 that shows the base object being inherited by all of the MorphBank objects. This information will be used later in Chapter 5 to show that the database can be proven to be correct and complete at any given moment if the above conditions are present.

4.2 Image and View Relationship

MorphBank requirements stated that images within MorphBank must conform to standards developed by the individual interest groups. The relationship between an Image and a View is therefore very strong. These standards are stored in the View relation, and although not show in Figure 4.2, each View is associated with a specific taxonomic range of biological entities. The related image must be of a Specimen that is also within the taxonomic range of the View. Standardized views will eventually allow scientists to develop tools that will assists biologists in the identification of species through image recognition software, a research topic under investigation in MorphBank at this time.

Any particular Image is restricted to an association with only one standardized View which in term has a specific taxonomic range. Conversely, many Images may

share the same View. One of the more important requirements to rise out of MorphBank was the desire of scientists to query the MorphBank information system on the images pertaining to a particular view and taxonomic range for the purpose of morphological studies. This feature will become more important later when discussing the requirements for phylogenetic character and state analysis.

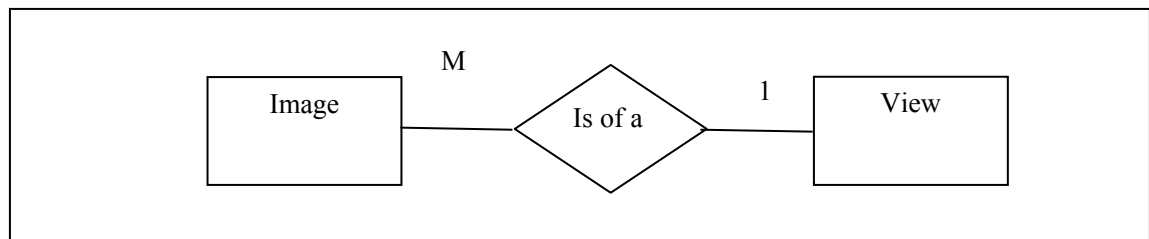


Figure 4-2 Image and View Relationship

4.3 Specimen, Image, and Locality Relationship

The Specimen relation is considered the pivotal object in MorphBank central to almost all other objects. Although only two relationships are shown in Figure 4-3, the Specimen relation references more tables through foreign key references than any other relation in the current version of MorphBank. The Specimen relation contains all of the information directly related to the specimen. During the course of requirements analysis, there was some controversy surrounding the formal definition of a specimen for the purpose of ontology. For instance, a specific plant, animal or organism is usually considered a specimen but so are the individual parts. limbs, leaves, hair, blood samples, stain smears, etc are all considered specimens as well as the entity from which they were taken. Another example came from an entomology laboratory which categorized an entire ant mound along with all of the colony's ants to be a specimen and then cataloged each individual ant gathered also as a specimen. The relationship Specimen also includes the ability to have an individual Specimen record be subordinate to another Specimen record in a separate relation.

Figure 4-3 depicts that basic Specimen, Image, Locality relationship. For the purpose of this discussion, only the references to the major objects of MorphBank are shown. Missing from this diagram are the association of the Specimen tuple with the

taxonomic name reference and other supporting tables such as those that describe gender, form, and other restricted characteristics. A Specimen may have associated with any particular tuple more than one image where as an image may be associated with only one Specimen. On the other hand, it is assumed that a specimen was gathered in a single locality. The original requirements of the system identified the possibility that the same locality would continue to be used for any specimen, regardless of its' taxonomic determination. However, this was not the case. As it turned out, the locality as well as the expedition and date when the even occurred was just as important as the location itself. Therefore, when the biologists traveled back to a location to gather additional specimens, a new locality record was entered into the database. Although the cardinality of the Locality to Specimen relation is 1-Many, there are instances where specimens gathered at the same physical location will have different Locality references.

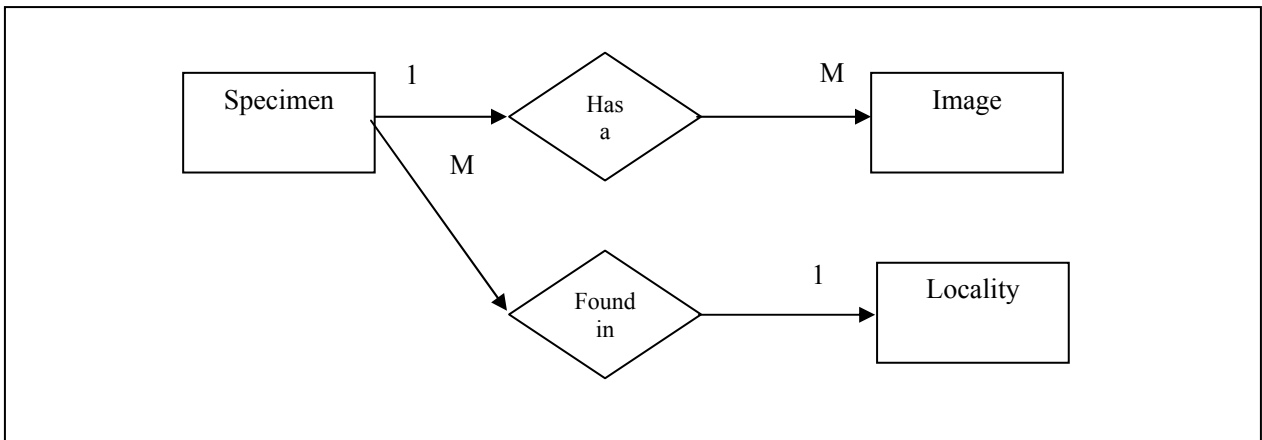


Figure 4-3 Specimen, Image and Locality Relationship

4.4 Collection Relationship

One of the major requirements for the new MorphBank system is the ability to group objects together inside the information system into private and public collections. A Collection will be defined as a set of MorphBank objects that share some type of user-defined relationship. The objects will normally share some type of relationship within the MorphBank system but this is not required for an object to be considered part of a Collection. Membership in a Collection is not restricted to the taxonomic determination of an entity either. Any major objects (Specimen, Image, View, Locality, Publication,

Annotation, or even another Collection) can be made part of a Collection. Because of the complexity of this particular relation (see Figure 4-4), specific rules were needed to ensure the integrity of the set. The rules are listed below:

1. A Collection must have at least one MorphBank object. Attempting to delete the last object of a Collection will cause the entire Collection to be deleted.
2. A Collection must have an order associated with the objects. Before the Collection is published, the owner may alter the order of the objects as they appear to users.
3. Since a Collection can be an object in another Collection, there is the concept of sub-collection.
4. A registered MorphBank user may create a new Collection from query results and before the Collection is made public (published) the owner may alter the order of the collection and the content by deleting or adding objects.
5. A registered MorphBank user may make a private copy of any existing published Collection or an unpublished Collection provided that User is a member of the group that owns that Collection. Once the copy is made, the user owns that copy and may make modifications.
6. Objects within the Collection will have names associated with them that are distinct inside the collection of objects. These names may be altered by the owner prior to publication of the Collection.
7. Objects within a Collection may be viewed and annotated similar to other objects within MorphBank.

Collection objects have the same characteristics as all objects within MorphBank in that they can be exposed through Life Science Identifiers, made available through web service calls, and through the MorphBank Show function. In terms of relationships, a MorphBank Collection object is by itself a single entity with one or more MorphBank objects associated. Only the unique MorphBank id, user id (creator), group id (group owner), and Collection Name are stored with the Collection object. In a separate relation, the individual object id numbers along with the object's order and object type (what relation id) it is. A Collection will have one or more associate objects and an object may belong to several Collections. Since a Collection can be an object of another Collection, the

system must check for circular relationships. Therefore the notion of a parent identification is introduced that pertains to only objects of type Collection. This single branch tree structure was developed to ensure that a Collection could not be a descendant of itself.

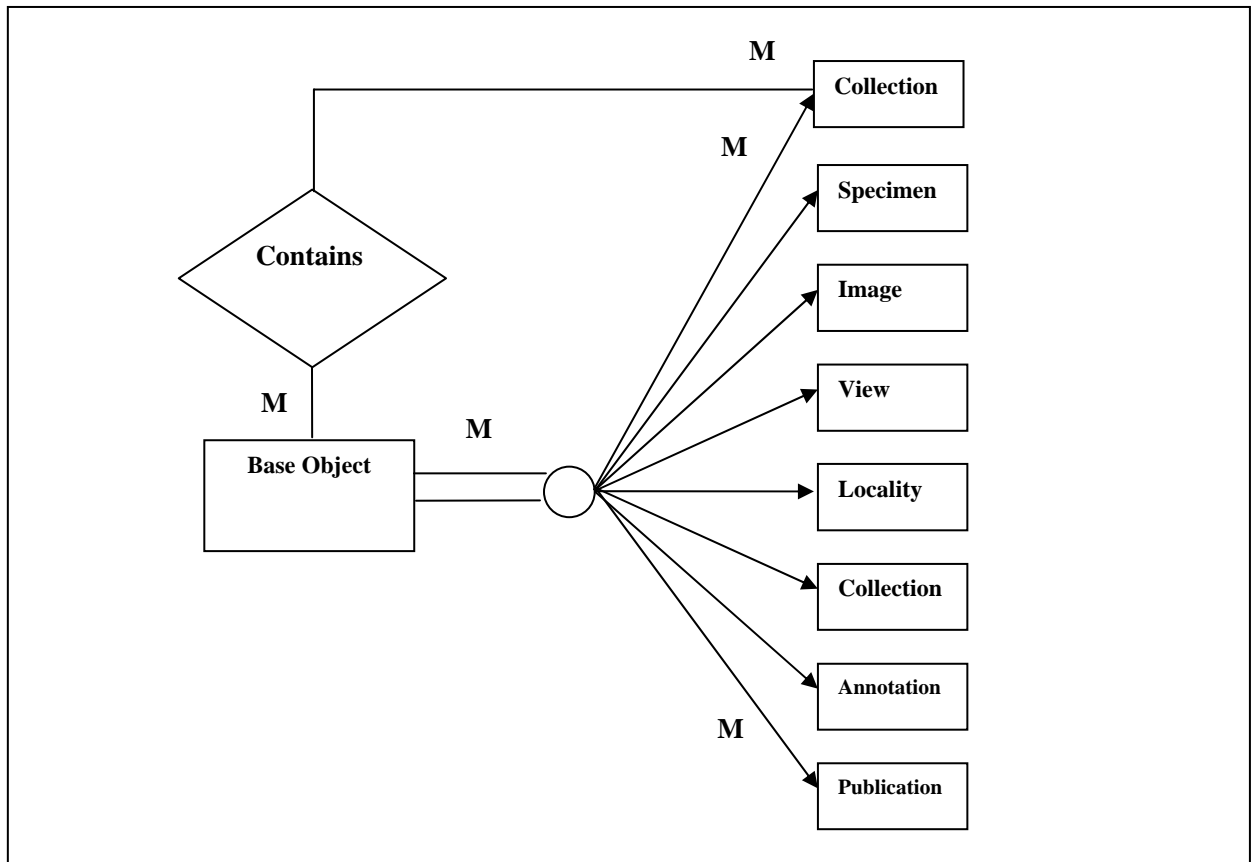


Figure 4-4: Collection Relationships

4.5 Related Objects

The Related Objects extends the idea of collections by allowing objects in MorphBank to have an unrestricted but identified relationship with any other object in MorphBank besides the relationships that are already defined. The Related Objects concept originated out of the idea by the MorphBank Research Team to allow for a standard method of communication of unrestrained relationships to users. Early in the requirements analysis phase of the project, it was realized that any particular organism on earth could have at least an indirect relationship with almost any other organism.

Likewise, the various object within MorphBank (Image, Specimen, View, Locality, Collection, Annotation, and Publication) could have an unlimited number of relationships with any other object. Rather than define these relationships through foreign key definition, the conceptual model uses the idea of inheritance of the data items within the Base Object to make a many-to-many relationship among the other object.

The following table 4-1 illustrates the concept of Related Objects. First, the relationship must be between two defined objects within MorphBank. Second, the relationship has direction. The direction can be left to right (The first object is related to the second), or duplex (the relationship definition is bidirectional such as in the case of siblings where a bother is related to another brother).

Table 4-2: Object-to-Object Relationships

<u>Left Object</u>	<u>Relationship</u>	<u>Right Object</u>	<u>Direction</u>
Specimen	Is a food source of	Specimen	Right
Specimen	Litter mate	Specimen	Duplex
Specimen	Identified in	Publication	Right
Collection	Defined in	Publication	Right
Specimen	Is a parasite of	Specimen	Right

The Related Object relationship is a many-to-many cardinality. A Specimen can be related to many other Specimens, Images, Views, etc and likewise. In a previous section, it was mentioned that a Specimen could be subordinate to another Specimen record. The Related Objects relation allows for this definition within MorphBank.

4.6 Annotation Conceptual Model

The requirements that helped create the conceptual model for MorphBank annotations evolved over several years originated from the concept of manual annotations on biological specimens. The desire to be able to annotate any object in MorphBank has never changed. However, the concept moved from the original idea of a graphical concept to one of ontology. Scientists deal in concepts. The notion of

assigning a string of taxonomic names to a range of entities is a concept in itself. The use of the scientific names conveys a meaning that is understood by the biologists and is part of their ontologies. Annotations are a means to extend the ontology by adding association to objects and thereby refining their meaning. Storing, efficiently locating, and displaying additional ad-hoc data in MorphBank is the prime motivation for annotations.

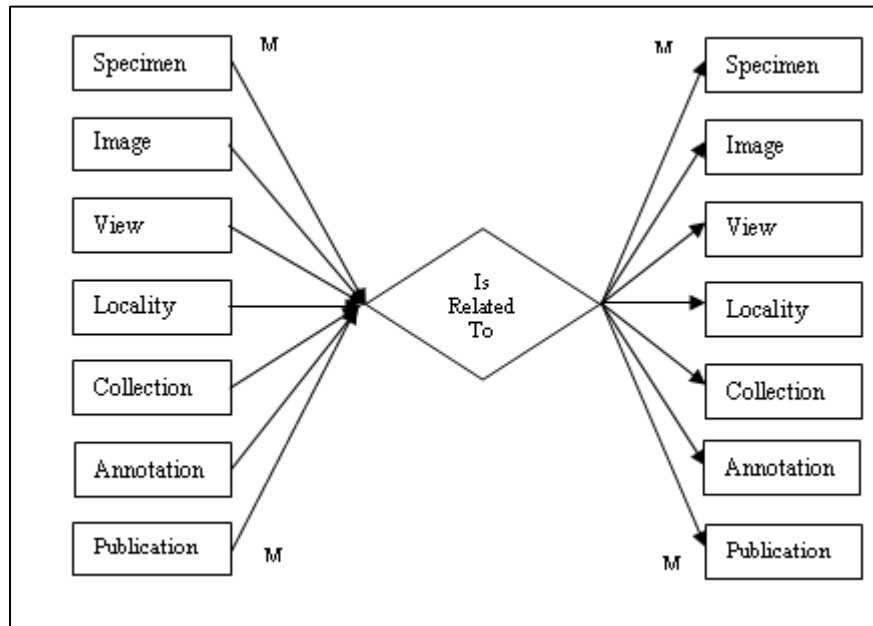


Figure 4-5: Multiple Relation Diagram

Figure 4-6 depicts the Annotation Model as designed for MorphBank. Notice the multiple relationships that exist between the Annotation relation and the MorphBank Base Object relation. An annotation itself inherits the Base Object data to include the person who created the annotation, the group under which the owner declared ownership, the date it was created, the release date of the annotation, and the date it was last modified. A MorphBank Annotation must itself annotate another Base Object but one annotation can only be associated with one Base Object at a time. However, since a

Collection can be annotated, an Annotation can annotate several items at one time.

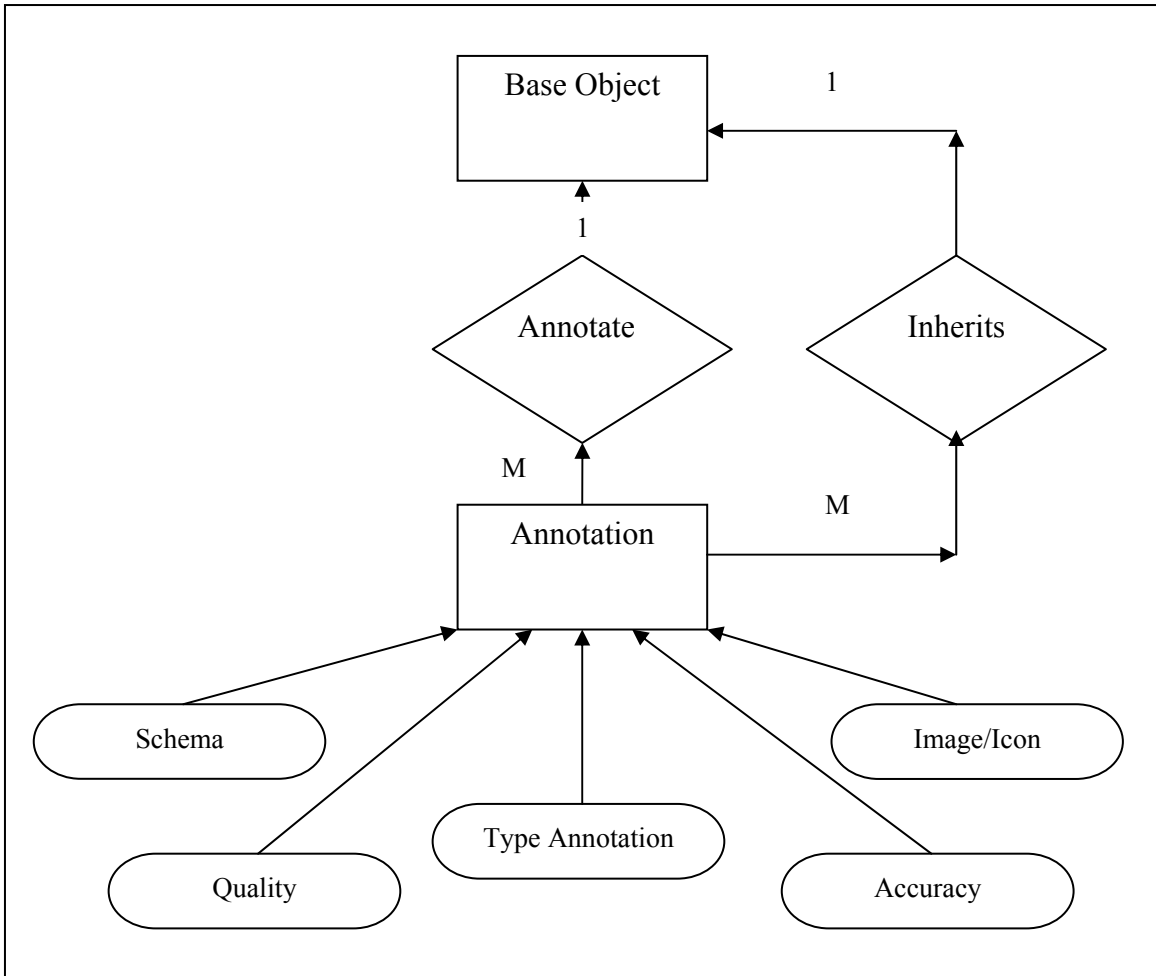


Figure 4-6: Annotation Conceptual Model

Each Annotation will contain a reference to either an Image or some type of icon that will permit the visual identification of the object being annotated. Each Annotation will have a type that can be one of the predefined types in MorphBank (General, Legacy, XML, or Determination). A General annotation is one where any ad-hoc comments are included with the annotation. Legacy annotations are those that were made either in another automated system or one physical specimens some time in the past. While the original author of the annotation may not be a MorphBank user, their annotations can be made available to other users. An XML annotation is the most flexible of the four types. Since MorphBank is incapable of internal schema modification it is important to

introduce the concept of an extensible schema. Through the XML type of annotation, we give the user the ability to define or select an existing XML schema in MorphBank and then upload an XML document that precisely defines that data in which the author wishes to associate with the annotation. Using this method, we can include highly organized and highly searchable data into MorphBank. The Schema along with the XML document is stored in MorphBank. Both the quality and accuracy of an annotation can be quantified. The quality of the annotation can be defined by the qualifications of the person who is performing the annotation, the recency of the annotation, the primary taxonomic rating of the person performing the annotation, and the number of individuals agreeing with the annotation. Finally, we introduce the idea of Determination Annotations. Indigenous to biology, determinations are simply the formal identification of specimens and the association of taxonomic names to that specimen. The complexity of this particular implementation will be discussed in Chapter 6.

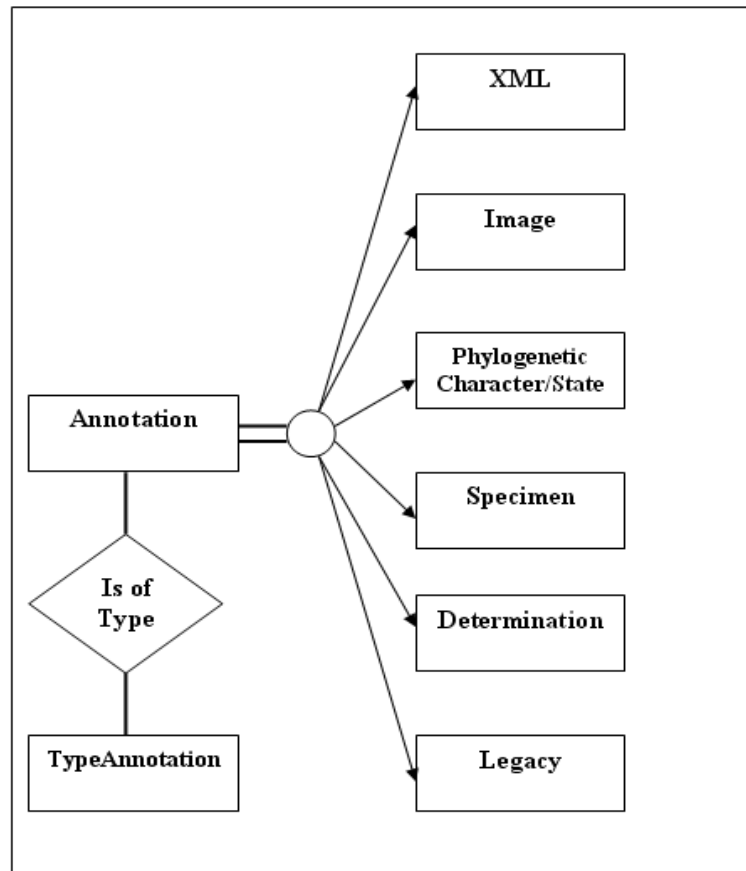


Figure 4-7: Annotation Inheritance Model

The display of annotation information is divided into six parts as illustrated by Figure 4-7. Part A displays the information about the object regardless of the type of annotations, Part B displays either the image (for a specimen or image) or the icon that represents the object. Part C contains the narrative portion of the annotation while part D contains all of the specific object data itself to include the inherited Base Object data. Part E is optional and only contains information if the type Annotation is either an XML extensible schema or a determination annotation. Part F contains all related annotations that belong to the object or annotations that are indirectly related through no more than one degree of separation.

4.7 Security Conceptual Model

Reliability and accuracy are considered main objectives of the MorphBank information system. Additionally, users want to be able to search for information in MorphBank based upon the reputation of the contributors and the accuracy/completeness of the data. The section will describe the design model intended to satisfy the security requirements and provide the relationships in the data to both assure that the data can be validated and can be ordered according to the rating of reliability. Users registered with MorphBank will have no more than one user account with the system. At the time of registration, the user will be assigned a privilege Taxonomic Serial Number (TSN), Primary TSN, and Alternative TSN. A description of these identifications are defined in Chapter 3. At this time the user's expertise will be rated on a scale of 1 – 5 by a panel of his/her peers with a 1 being an amateur biologist with an interest in morphology through a 5 being a world renowned expert.

A unique username will be assigned by the MorphBank Administration and will be given an initial password. They will have the authority to change the password and personal information through the personal user maintenance account. All users of MorphBank, regardless of the version, will have the same basic privileges in the system to browse and search images on MorphBank (see figure 4-8 for Use Case scenarios). Additional access is granted by membership in groups managed by a Group Coordinator. Groups are identified by a name, the taxonomic range for which they responsibility, and the MorphBank objects for which they own. The taxonomic range is identified by the Taxonomic Serial Number (TSN) as used in the ITIS database system. For instance, a group responsible for images and data in the family Hymenoptera would have a TSN

assigned as 152741 and would allow users of that group access to related objects with that TSN range or lower. Users may be members of several groups. If for instance, a user is a member of the Hymenoptera group and also of the Cephus cencus (Wheat Stem Sawfly) they would have access rights to objects owned either group. However, to maintain the integrity of the data, they would have to declare the group they belong to when modifying or adding data in order to maintain data ownership.

Annotation Record [77439] Title: Specimen	
Base Object Data And Type Annotation A	Image Primary Specimen Image Or Icon B
Comments or other related text information C	Specimen, Image, View, Locality, Annotation, Collection, or Publication data D
Determination Annotation Data Or XML Data E	Show related Annotations and Related annotations With one degree of Separation F

Figure 4-8: Annotation Display Model

Object Ownership: MorphBank will have the ability to track ownership of individual objects within the database. Ownership is comprised of (1) the person in MorphBank that contributes the object, and (2) the group that will have primary responsibility for verification and validation of the data. While the object is under review (prior to the release date) MorphBank users that do not belong to the object's group may not view the data, modify it, nor annotate it. While the object and associated data is being entered and no release date established, only the owner (contributor) of the object may modify it. The Coordinator of the Group and Lead Scientists who are members of the group; may view or annotate the object. Once an item has been released, all users (including the world) may view the object and all MorphBank registered users may annotate the data.

Rating of Users: Users are placed in categories that depend upon their area of expertise. To determine the range of responsibilities of a person there are four associated Taxonomic Serial Numbers (TSNs) associated with each user. They are as follows:

5. **Privilege TSN:** This TSN value identifies the highest class of biological entity for which the User has education, training, and experience. When requesting a MorphBank User account, a user will state their highest taxonomic range. For example, a person whose recent research involves wasps, ants, and bees may request their Privilege TSN to be 99208(Class: Insecta).
6. **Alternative Privilege TSN:** This TSN is usually a value in another branch of the Taxonomy Tree that would allow a user to hold scientist or lead scientist privileges in another group. For instance, specialists in the area of plants may have a detailed knowledge of bacteria that affect plants in their specific area.
7. **Primary TSN:** The Primary TSN identifies the user's specific area of expertise. This TSN usually indicates that the user has done extensive research in this area and is considered by his peers as an expert over this range. This TSN range is usually at or below the family level. As an example, an individual with extensive experience in certain parasitic wasps may have a Primary TSN of 150425 (family: Cynipoidea).
8. **Secondary TSN:** Often, users will belong to working groups that have a wider taxonomic range than their own specific expertise. Following the previous example, a person whom does research on Cynipoideas may belong to a

Hymenoptera working group and have a Secondary TSN of 152741 (Order: Hymenoptera).

When a user requests access to a MorphBank object, there are several standard steps that must be accomplished before the request can be completed. First of all, users that are not logged onto the system are not given access to the screens that permit the addition or alteration of MorphBank data. Users with accounts that have logged onto the system and declared a group they are working under can request access to MorphBank objects. The first step is to determine if an object has been published. Released data is visible to the world and cannot be altered. Only unpublished data can be altered. Second, the creator and group ownership are determined. If the user requesting data does not either belong to a group or is working under the auspices of another group different from the object, he/she are not given access. Finally, the user's taxonomic privileges are checked to ensure that they have responsibility for the taxonomic range requested to modify, add, or annotate data. Later in this chapter, a concept to protect MorphBank data through web service access will be discussed.

4.8 Life Science Identifiers Model

Scientists associated with MorphBank required that information stored in the information system be made readily available to other scientists around the world. As such, each identifiable bit of information in MorphBank must be uniquely identifiable and persistent. The MorphBank server has moved four times within the last two years which has caused some problems. There are multiple copies of the database and there have been times where the locally unique keys have changed. This presents a problem to scientists who are attempting to use systems like MorphBank as permanent references for their published research. Early in the program, it was decided to adopt a Globally Unique Identifier strategy for objects in MorphBank that would have an external exposure requirement.

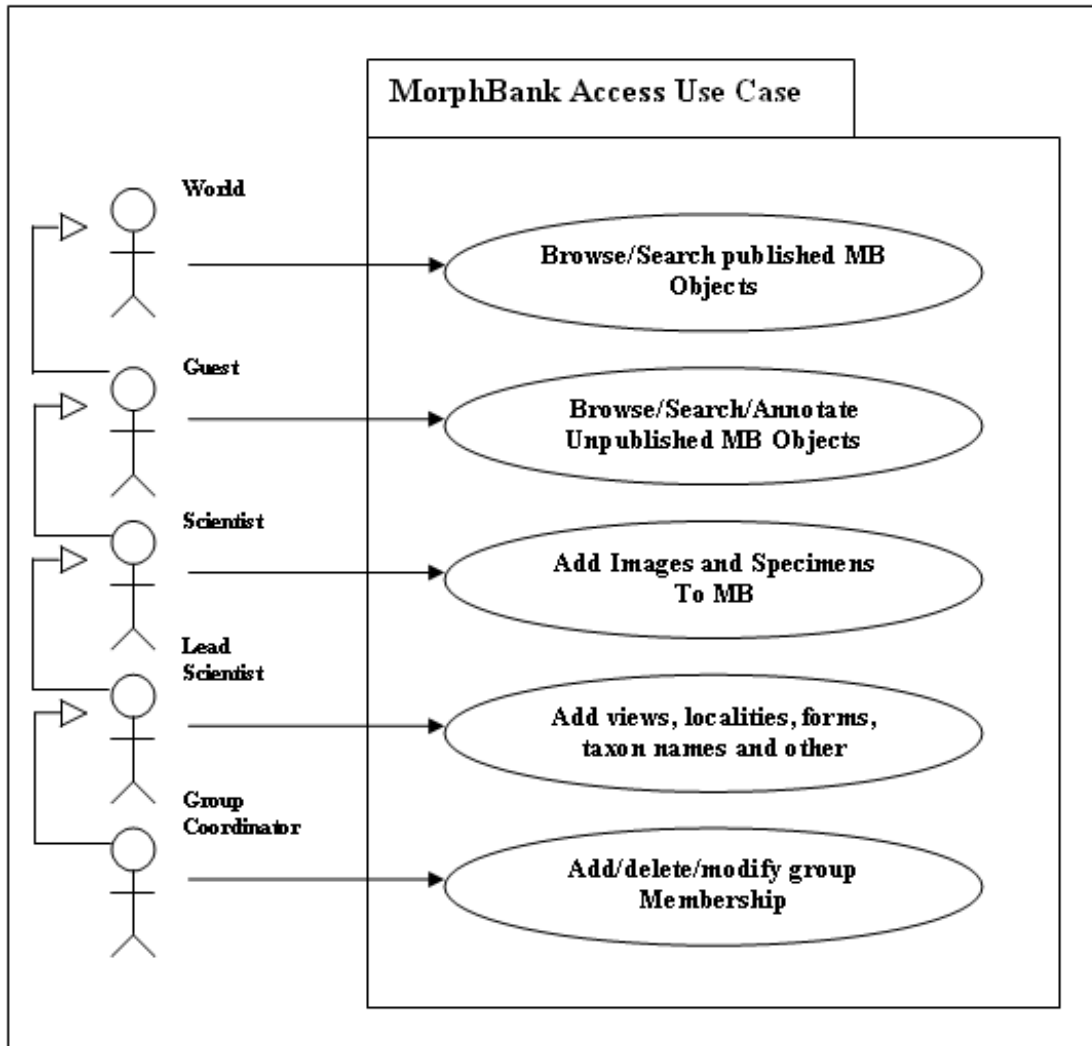


Figure 4-9: MorphBank User Privileges Use Cases

To solve the problem of consistency, the designers of MorphBank decided early in the project conception phase to select LSIDs as the Globally Unique Identifier for which will be used to expose objects to the world. Life Science Identifiers are defined as follows: “The Life Sciences Identifier (LSID) is an I3C and OMG Life Sciences Research (LSR) Uniform Resource Name (URN) specification in progress. The LSID concept introduces a straightforward approach to naming and identifying data resources stored in multiple, distributed data stores in a manner that overcomes the limitations of naming schemes in use today. Almost every public, internal, or department-level data store today has its own way of naming individual data resources, making integration

between different data sources a tedious, never-ending chore for informatics developers and researchers. By defining a simple, common way to identify and access biologically significant data, whether that data is stored in files, relational databases, in applications, or in internal or public data sources, LSID provides a naming standard underpinning for wide-area science and interoperability.” [<http://lsid.sourceforge.net>]

LSID is a naming standard for distributed biologically significant data including files, database records, and data objects. The data should be accessible over public or private networks owned by a variety of academic, research, commercial, or government agencies. As a globally unique identifier, LSIDs are not semantically connected to the objects they identify. Therefore, no information about the object can be obtained directly from the identifier. LSIDs replace the physical addresses normally used to gain access to data objects. An LSID resolver returns a Web Service Definition Library (WSDL) file that describes the methods used to retrieve the data or the query needed to obtain additional data. Using LSIDs, MorphBank will be able to distribute the stored data among several physical resources without the need to keep and maintain their exact physical address.

LSIDs have five parts: the Network Identifier (NID); the root DNS name of the issuing authority; the namespace chosen by the issuing authority; the object id unique in that namespace; and finally an optional revision id for storing versioning information. Each part is separated by a colon to make LSIDs easy to parse. The main requirement for a piece of data to be considered an LSID is that it must be unique and persistent. In order to satisfy the requirement to be LSID compliant, MorphBank required that each individual object within the database be uniquely identifiable. An example of a MorphBank LSID is shown in Figure 4-9.

4.9 Web Services

The term web service is used as a broad definition for functions and services provided over the World Wide Web [<http://www.w3.org/2001/01/WSWS>]. All of the large industry leaders such as Microsoft, IBM, Apple, and Sun have services strategies. Web services references the architecture, standards, technology and business models that make remote access to data possible. In this section, we will give a brief overview of web services and then describe an overall strategy for implementing a web services

```
urn:lsid:morphbank.net:323091
    This object references an image within MorphBank. The
    actual location of the image may or may not be located
    at the server referenced by the address.
urn:lsid:morphbank.net:3339087
    This object references the specimen associated
    with image in the previous LSID.
```

Figure 4-10: LSID MorphBank Examples

Since MorphBank was envisioned as a distributed digital library, the web services paradigm is easily applied. The digital library community is used to applying verbs similar to those used in the description of biological entities for searching information. The web services architecture applies these verbs to search queries on database systems. Web services takes the requests, verifies the authenticity of the requests (and user), distributes the requests to the appropriate module, collects the results and returns the data back to the user.

IBM has published its web services architecture [<http://www.ariadne.ac.uk/issue29/gardner/>] which captures the infrastructure required to support web services in terms of three roles - service provider, service requestor and service registry - and the verbs describing the iterations between them: publish, find, bind (Figure 4-10). Bind is the step that allows an application to connect to a web service at a particular web location and start interacting with it. The basic IBM web services architecture is show in figure 4-10.

The MorphBank Web Services architecture is described in Figure 4-11. A user makes a request to the MorphBank system through a standard web services using a standard WSDL document in XML format. The MorphBank Web Services Interface takes the request and deciphers the XML document. If a user is making a request to modify the information contained in MorphBank or requesting unpublished data, a call is made the Security Module which takes in the verifies access through standard WC3 standards using MorphBank access data. The request for data or modification is then

sent to the Data Request Resolver which determines what portions of MorphBank are involved.

Should a user request data from the MorphBank database, the standard routines for those modules are formed and called. If a request to modify the database is made, a simple acknowledgement of success or failure is sent back to the user via the MorphBank Web Services Interface. A request for data will cause a query to be formed and the resulting dataset to be returned to the MorphBank Web Services Interface where the data will be transformed into a WSDL data file and returned to the user. Formats of these documents have not been finalized yet although a brief experiment on a simple web services has been accomplished.

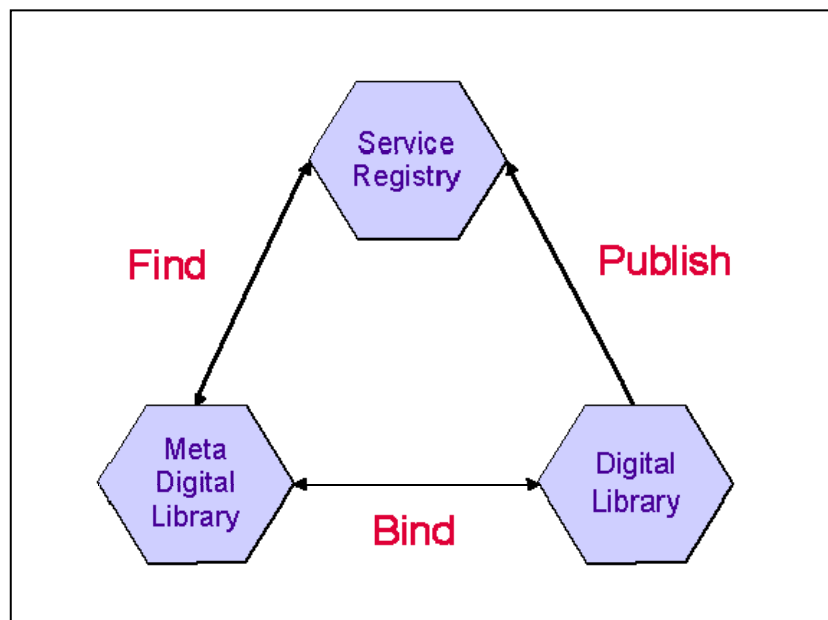


Figure 4-11: IBM Web Services Architecture

One of the early requirements of MorphBank was to create an interface with a collaborative project called MorphoBank. “MorphoBank is an online database and workspace for evolutionary research, specifically systematics (the science of determining the evolutionary relationships among species). MorphoBank is project-based; meaning a team of researchers can create a project and share the images and associated data exclusively with each other. When a paper associated with the project is published, the

research team can make their data permanently available for view on MorphoBank where it is now archived.

“The phylogenetic matrix aspect of MorphoBank is designed to aid systematists working alone or in teams to build large phylogenetic trees using morphology (anatomy, histology, neurology, or any aspect of the phenotype) or a combination of morphology and molecular data. In contemporary systematic methods in which morphology is used to build trees of species, one starts by constructing a matrix made of characters and taxa. Characters are features of an organism that appear in different forms. MorphoBank version 1 was funded by National Science Foundation grant DEB-9903964 to Maureen O'Leary and with financial assistance from the American Museum of Natural History. MorphoBank version 2 was funded by NOAA (NA04OAR4700191) [<http://www.morphobank.org/>]”

Since Morphobank maintains primarily phylogenetic character data, access to a large image repository became important. MorphBank developers created a simple means for Morphobank administrators to associate their data with MorphBank images using the show function. A fully functional LSID implementation was not ready to allow users to discover data. MorphBank administrators created a simple web service call that allows users to request a simple query on MorphBank data (see figure 4-11) and have returned a simple XML document that describes the images, the identification numbers, and the corresponding LSID. Morphobank or other users can dissect the XML document and store the URLs of the Show function for access to MorphBank images. At this time only images are searched and only published objects. Future versions will allow access to all published MorphBank objects.

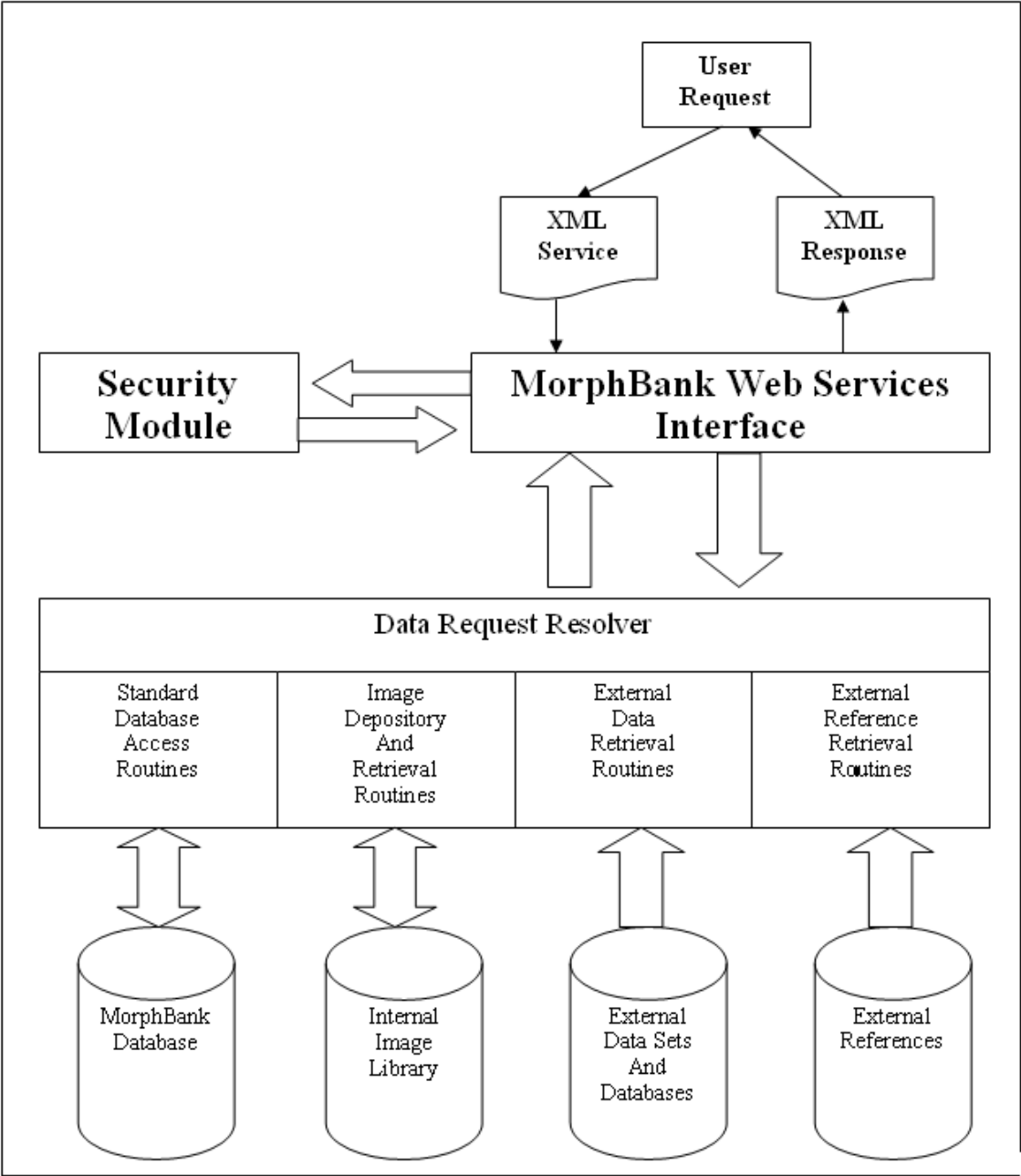


Figure 4-12: MorphBank Web Services

CHAPTER 5

SEMANTIC ANNOTATION

This chapter 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 Data Grid technologies to facilitate the capture, organization, and presentation of biodiversity information. 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, because XML documents can be indexed and searched using their associated schemas.

The discovery, identification, and documentation of biological entities are 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. However, with the increase in the discoveries of new organisms and a decrease in number of senior specialists, identification and curation of data have become more difficult. Often, collaboration required scientists to travel to the location of the specimens or for specimens to be sent to the scientists for first hand examination.

There are several problems associated with the discovery of information in biodiversity systems. The first is finding an image associated with a specific species and genus, finding information about that image and its association with other images, and the second is finding ad-hoc data about the images entered by biologists. Discovering ad-hoc data is the most problematic. As long as the data is well formatted and constrained to the database schema then finding and retrieving data is relatively simple. However, as was discovered, there is no practical limit to the amount of information that a scientist may wish to store for a particular specimen.

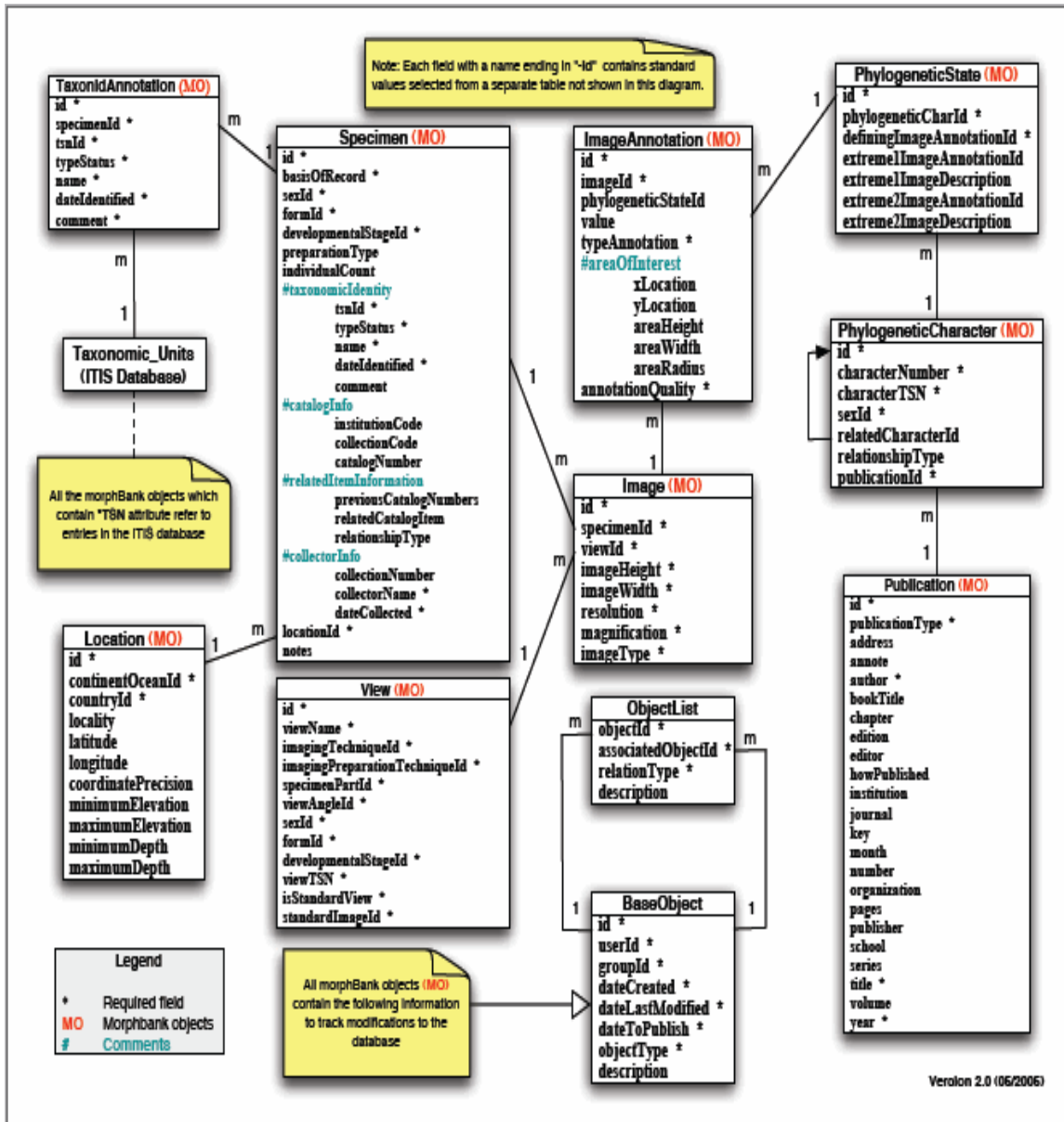


Figure 5-1: MorphBank Schema

In this chapter, I present a method using existing technologies that allow scientists to use their own schemas in describing and storing information. The approach combines the advantages of a highly organized relational database and extensible XML schemas to provide a flexible framework for efficient semantic queries.

MorphBank currently contains about over 60,000 images that are publicly available and approximately 250,000 images are expected to be released during the Spring of 2007. These images document a wide range of organisms, from plants to insects. A major

advantage of MorphBank is that images and associated data are maintained in a system based upon open standards and free software that facilitate the development of tools for image uploading, retrieval, annotation, collaboration, and other related tasks (see Figure 5-1).

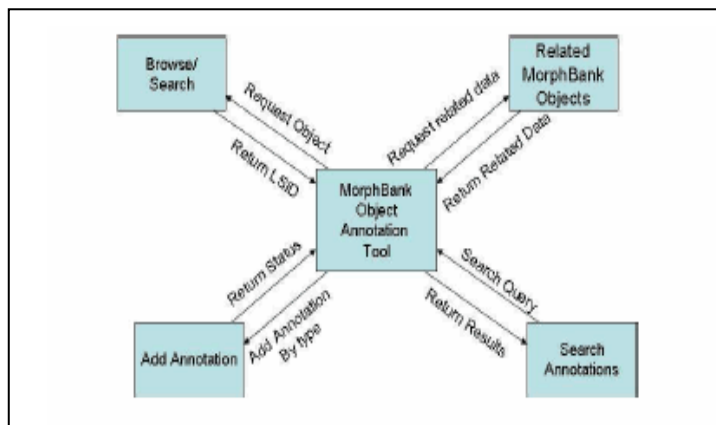


Figure 5-2: MorphBank Annotation Architecture

There were numerous problems with current biological databases that hampered the ability of research biologists to store and retrieve information. Each scientist maintains their own unique method for organizing and storing information. There have been several attempts to create client based graphical annotation tools. These annotation tools are usually client based and share no direct interaction with a centralized data repository. Therefore, the user must search and extract the data onto the client before the data can be annotated. Once annotated, the data must then be sent back to the repository. The annotation tools are usually highly specialized for a small set of users or so general that it provides no significant advantage. I developed a new approach to image annotations that addresses each of these deficiencies.

The original MorphBank Database schema did not incorporate any ontology standards currently recognized in the industry [SpMeJa03]. The MorphBank Research team selected the Darwin Core Standard [Meng04] for data item names and type representation. Although most biological ontology standards are still relatively new the Darwin Core Standard is the most complete set for biodiversity data. Most scientists today use the World Wide Web to share data and to search for information pertinent to their own research. To accommodate these standards the MorphBank system was

designed to take advantages of the innovations in web services and grid technologies. Additionally, flexible and extensible annotation requirements were built into the design.

The MorphBank Research team decided to organize the information within the database in an object oriented manner with the assumption that other interested groups would eventually access the data through web services. The database tables were logically organized along with their related member functions into seven logical services: Image, Specimen, View, Annotation, Geographic Location, Publication, and Collection (See Figure 5-1). Other supporting and security information or objects may not be visible through web services.

5.1 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) [SmMaSz05], the location of the object (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 information to easily find and catalog the data and associate related items without implementing a varied number of unique keys with different data types. This service creates new objects, catalogs them, and is responsible for maintaining database integrity and consistency. The use of LSIDs allows MorphBank to grow without radically altering the underlying database structure by allowing any object that can be identified via an LSID to be incorporated into the data and annotated.

Since each MorphBank object is identified using LSIDs, the use of foreign keys within the database is not restricted to a single table. For instance, an Annotation object may be associated with an image, specimen, location, user, group, or even another annotation (see Figure 5-1). This 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 pre-defined relationships in MorphBank, the use of LSIDs allow users to define an unrestricted set of complex relationships of objects within the confines of the system.

5.2 Biological Annotation

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 [Hala01]. A major restriction placed on the development was that 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 data. Updates and annotations made by one scientist must be readily available to other colleagues for collaboration in a timely manner.

There has been considerable effort put into the development of a general purpose web-based annotation toolsets over the past several years. In their paper on web annotations, Venu Vasudevan and Mark Palmer [VaPa99] described in 2000 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 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 [Marsh97].

The basic 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. Through the use of MorphBank and a web based annotation tool, many of these problems were solved.

5.2.1 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 Co laboratories under the guidance of Dr. Jim Myers [MCGS04]. “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 [Myers04].” However, many of the current set of applications under development assume the underlying schema and functional applications are highly developed using well established industry standards and organized for access by web services applications. Database applications developed using an evolutionary approach to design require the development of wrapper software in order to take advantage of these products [Hass96]. These wrapper products translate the ontology of the database into a common definition using W3C standards such as an RDF Schema approach.

MorphBank was designed to allow users to take advantage these middleware products by conforming to industry practices and standards while maintaining the ontology of the original data. Figure 5-2 depicts the MorphBank Annotation Architecture. Users can browse or search the web site for MorphBank objects using a variety of tools provided through the web site. As the thumbnail or object is viewed, users with appropriate privileges can annotate that object. The tool loads the selected object and automatically queries the database for any related data. The object is displayed along with the first annotation (if any exist). All other related annotations are shown in a scroll bar and can be selected for viewing. Also, there is currently a simple set of query tools that allows users to search annotations for related images, related species, and by image view sorted by contributor, title, or date. Users are free, at any time, to add their own annotations. Annotations are immediately stored in the database and can be searched at once by other users with appropriate permissions. The images and associated annotations are also available to users connecting to MorphBank using web services.

5.2.2 The Purpose of Annotations

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 [KoMaSc05]. 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. In general, an annotation tool must be simple to use, fault tolerant, precise,

allow for general comments, and permit the users to place annotations that have specific meaning.

Annotations are a way of adding additional information in the database without altering the schema of a system or the original image. Adding XML content to annotations increase the utility of the system by extending the number of type of data items that can be stored. The other problem is of information discovery. There are numerous techniques available that would perform ontology searches on plain text data that are normally stored in on-line annotations. However, the results usually represent a guess as to the requestors meaning. By storing the annotations using an XML schema, exact meanings are associated with the data removing any ambiguity. The results of these queries return only the records that correspond exactly to the request of the research biologists.

5.3 Association Annotations

The need to annotate images and other objects in MorphBank is recognized as a major requirement for scientific collaboration. To increase the efficiency of the discovery of information, annotations have been grouped into categories that each has particular data and functional requirements as well as heuristics.

- **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 the 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 identity of the species and genus of a specimen. Users select a specimen and use the associated images to make a recommendation as to the specific genus and species identification. The identification is connected with a Taxonomic Serial Number (TSN) in the Integrated Taxonomic Information System maintained by the United States Department of Agriculture (USDA).
- **Phylogenetic Character and State:** This type of annotation is 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 selects from the database a genre of phylogenetic characters and a particular character and state. Similar to the location annotation, the difference is that the location of the annotation on the image is associated with a particular record in the Phylogenetic Character-State database.
- **Relationship:** There are already pre-defined relationships built into the MorphBank database that were defined as part of the original requirements. Such relationships include the ability to associate a specimen with multiple images. Also, images can be grouped by standardized views, collector and location. Relationship annotations allow the user to define additional relationships. User can 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 predefined 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 added benefits of decreasing the search time for annotations and reliability of the information.

A Specimen image annotation captures 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 group of specimens in a single transaction. Image annotation associates textual information to the specific region on an image to enable semantic querying. Two technologies are frequently used: Text-based approach and field-based approach. The former simply added 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 5.3 shows a simple image annotation of the field-based approach.

MorphBank has the ability to add multiple annotations to a single image without modification of the original image. Each annotation is a separate instance that can be queried and associated with the image, specimen, locality of the specimen, view of the image, user who added the image, and all other related objects. Not only does an image annotation capture knowledge of related species of the image but it also creates relationships with other MorphBank objects that have some link with the image. Additionally, by storing the annotations apart from the image, MorphBank preserves the original image as a specimen would in a museum and the number of annotations has no practical limits.

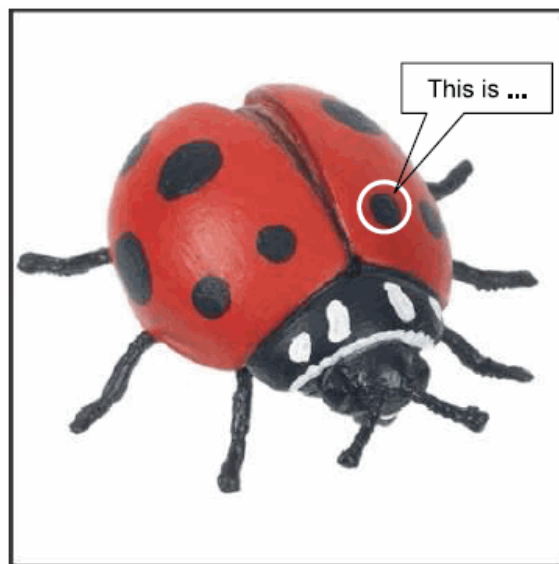


Figure 5-3: A Simple Image Annotation Example

5.4 Image Annotation

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 with text is not suitable for the more sophisticated requirements. The heterogeneous data models from different biologists and the diversity of biodiversity require frequent update and different data structures in the information system. Creating dynamic tables in relational databases for different data models is not practical in MorphBank while taking integration constraints into consideration. One of the original requirements for the system was to use a relational database to store the textual information and a basic file structure to store the images. Using this architecture, the names of the attributes are relatively static. However many biologists use a variety of naming and organizational conventions, many of which do not conform to any known specification or standard. So the question is how to take advantage of the power of a relational database but give the scientists the flexibility they require.

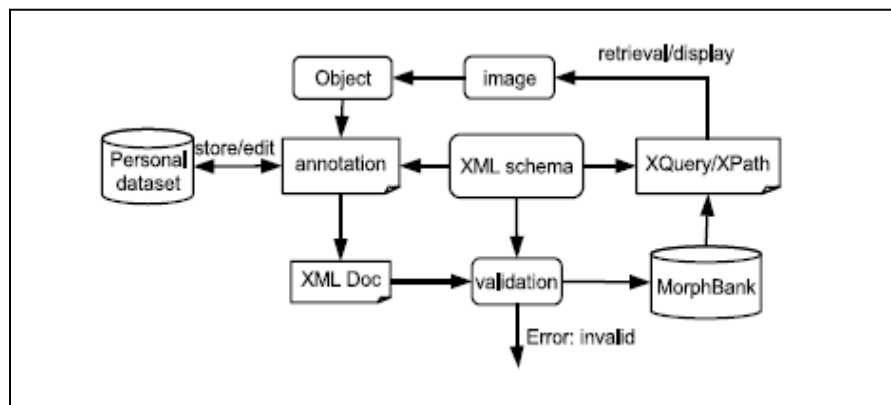


Figure 5-4: Image Annotation Overview

MorphBank stores some of the annotation information in a relational database using a W3C XML document [www.w3.org/XML/]. This structured, human readable and self-description XML document can be easily parsed using standard XML parser and can be easily extended. An XML document has several distinct advantages. Firstly, it provides the biologists with the flexibility to annotate an image using their own data model. Secondly, it improves retrieval performance by combining relational database

technology and XML querying such as XQuery [www.xquery.com/] or XPath [www.w3.org/TR/xpath/]. Thirdly, an XML document can be easily validated by W3C XML schema [www.w3.org/TR/schema]. Fourth, the association of a data item with an XML field is exact and is not open to interpretation.

5.4.1 Image Annotation Architecture

Figure 5-4 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. Users can choose a schema from a set of predefined schemas or create a new schema. Annotations can be saved in progress to be edited further, or can be imported into the MorphBank system. The information of an annotation is stored in an XML document along with an instance of the XML schema which imposes structural and semantic constraints on the data. 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 subset of the XML document.

5.4.2. Schema Customization

MorphBank image annotation tool has a set of predefined schemas. It also allows users to create their own XML schema. User defined schema extends the predefined schemas to meet individuals needs. All of the XML schemas are derived from a generalized schema, which defines the information of an image. Attachment D contains an example of a complete example of the schema for an image annotation in MorphBank.

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 this general information. For example, Figure 5.5 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 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 5-5: An Example XML Annotation Document

5.4.3 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 enumerations 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.

5.4.4 Query 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 subset 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 [www.w3.org/XML]. The query interface automatically enumerates the optimized querying to improve the query performance.

PLANT PATHOLOGY HERBARIUM OF CORNELL UNIVERSITY
ANNOTATION SLIP: Specimen No. CuPiCH 11
This is apparently an undetermined specimen of polydesmia.
Probably close to *Psilocybe Dumontii*.
Annotated by: Richard I Koff Date: 13 February 1983
 Slide
 portion (*check box or boxes when applicable*)
 _____ Kept in Herb. _____ as No. _____

Figure 5-6: Sample Herbarium Annotation

5.5 Annotation Integrity

Not all annotations are created equal. With the ability to store virtually an unlimited amount of information, finding the information and ranking the data in accordance with accuracy and relevance can be a problem. In this section, I describe a strategy built into the design of MorphBank that addresses these concerns. Using only existing data and the security module, Data can be searched and ranked based upon a non-subjective rating scheme. Remember that the roles of a user are stored in the MorphBank system (guest, scientist, lead scientist, coordinator, or administrator), the taxonomic expertise of a person, and the relative qualification of a person (1-5). Additionally, the number of times an annotation was referenced and the number of times a determination annotation was agreed with or disagreed is tracked. The reliability of the specimen and image can be used in this calculation especially if a specimen represents some type of type status. Using this data can increase the efficiency of retrieved data.

5.5.1 User's Role Weight

When a user enters data or annotations they must also login to one of their groups to which they have membership. The higher the role a person holds within a group the more reliable the data. Additionally, a person's expertise can increase over time and data and annotations entered later in their career are considered more valuable and therefore more reliable than the data they entered previously. Annotations and data entered would retain knowledge of the role of the user processed at the time of data entry. Using a simple association of assignment of a weighting factor with a user's role, a rating schema can be incorporated. Since anyone with only world access has read-only privileges, no rating will be assigned to them. A person with guest privileges has limited access to the data and has not authority to enter data other than annotations. Individuals with Administrative privileges will also be given no weight (rating=0) because they inherently do not have any biological expertise unless they have group membership as a biologist. Guest will be weighted as 1, Scientist as 2, Lead Scientists as 3, and Coordinator as 4.

5.5.2 Taxonomic Privilege Weight.

Users are assigned a Taxonomic Privilege based upon their resume and biological expertise. Designed to illustrate the user's breadth of expertise, this rating is usually assigned at the Order within the tree of life. For instance, a person conducting research in the area of parasitic wasps may have a Taxonomic Privilege rated at Hymenoptera. Subsequently, a weighting factor can be assigned to data on someone's Taxonomic Privilege. Certainly, someone entering annotations on specimens contained within the taxonomic tree associated with their Taxonomic Privilege would have a higher rating than someone who's Taxonomic Privilege was located in a different branch of the Tree of Life. The weight is based upon two factors: (1) Is the MorphBank object in the same branch of the tree as the Taxonomic Privilege and (2) How high in the tree is the Taxonomic Privilege. The rationale is that the higher in the taxonomic tree a person's rating is the more trusted they are. Example: Someone with a Taxonomic Privilege of *Drosophila Melanogaster* (Fruit Fly) would have a lower rating than someone with a Taxonomic Privilege of subphylum Hexapoda.

Table 5-1: Animal Kingdom Taxonomic Rank Id

Taxonomic Unit	Rank Id
Kingdom	10
SubKingdom	20
Phylum	30
Subphylum	40
Superclass	50
Class	60
Subclass	70
Infraclass	80
Superorder	90
Order	100
Suborder	110
Infraorder	120
Superfamily	130
Family	140
Subfamily	150
Tribe	160
Subtribe	170
Genus	180
Subgenus	190
Species	220
Subspecies	230

Each Taxonomic name has a rank id and a kingdom id. MorphBank can detect instantly if any object is in the same Kingdom. The rank id of a specimen gets higher the further down in the taxonomic tree of the taxonomic unit. Table 5.1 shows the taxonomic unit types and rank ids for the Animal Kingdom.

Someone with a Taxonomic Privilege of the Animal Kingdom should have the highest rating and conversely, a person with a Taxonomic Privilege of Subspecies in the Animal Kingdom should have the lowest. Simply take the rank id of the Taxonomic Privilege and subtract it from 240. This yields a higher number the higher in the Tree of Life. Example, a person with Family rating in the Animal Kingdom and still within the branch of the tree of the MorphBank object would yield a rating of $(240 - 140)$ one hundred. This would be higher than someone with a rating species which would only yield a twenty. A Taxonomic Privilege in another part of the Tree of Life only yields a rating of 10.

5.5.3 Primary Taxonomic Rating

A person's Primary Taxonomic rating works just the opposite of the Privilege Taxon. This identifies a person's specific expertise and the closer a Primary Taxonomic Rating is to the MorphBank object's Taxonomic identification the higher the rating. However, a person's expertise may or may not be in the same branch of the Tree of Life as that of the MorphBank object but may still exist within the user's Taxonomic Privilege. For this reason the lowest common point in the tree must be found that is shared by both the MorphBank object and the users Primary Taxonomic rating. For example, assume a person is annotating a specimen of a *Cynips douglasi* (parasitic wasp) however, the persons Primary Taxonomic rating was in *Xiphydria pilongata* (wood wasps). The common root would be Hymenoptera and a validity rating of 100 (order) would be assigned.

5.5.4 Other Rating Factors

Other factors considered in determining the rating of a MorphBank object are quantitative values within the system such as the age of the specimen, the number of positive determination annotations associated with it, if the specimen or object has a type status, and if the objects are used as phylogenetic character references. Since these values are slightly more subjective, the user is allowed to assign different weights to each one in order to customize the results of the query.

5.6 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 MorphBank version 2.5 for the determination of herbarium specimens. The team took a standard herbarium annotation card (see Figure 5.6) 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 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 validate any XML document.

Users making determination annotations on herbarium collections would then select **Determination** as the type of annotation. 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 5-7. 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 curation of specimens. Users of MorphBank were able to store their images and corresponding data into MorphBank using the standardized Darwin Core type dataset. Additional information they wish to include was placed in the system as user-defined annotations.

```
<annotation id = '12345' type='Taxon Identification -  
<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 5-7: Herbarium Taxonomic Determination

An XML document can be submitted to MorphBank as a validated annotation schema. Once accepted into the database, scientists can then import their data directly into MorphBank without altering the baseline schema.

5.7 Conclusion

I have described an existing problem in the biology community for storing and retrieving digital image information on biological specimens. 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. At the functional level, many users have developed their own proprietary solution to this problem. The results of this research show that this method allows users of MorphBank to have a well designed centralized digital image database with the flexibility of a privately owned collection. The work performed under the NSF grant by the MorphBank project provides the Tree-of-Life initiative with a stable digital image database and annotation tool set that is currently used by biologists around the world.

CHAPTER 6.0

ANNOTATION TRIALS

The major challenge of the research project was the magnitude of work that must be accomplished to prove the results. In order to gather sufficient information to formulate a conclusion, a trial of the prototype software was conducted during the Fall term of 2006 and continues as MorphBank is a fully supported interactive data based web site. MorphBank version 2.5 was released July 29th, 2006 in time for use in a remote annotation trial for a herbarium collection organized by Dr. Austin Mast and a similar annotation trial for hymenoptera specimens is planned by Dr. Fredrik Ronquist in the near future. The annotation software is connected to a fully functional phylogenetic biological image database currently under maintenance. This chapter describes the results of those trials.

6.1 Annotation Trial Objectives

There were several objectives to the trials some of which are related to this research and others related to the specific objects of the Principle Investigators of the NSP Grant that supports the research. As of the writing of this document, there does not exist another fully functional, complete, documented, and quality controlled database that research biologists can use for the repository of images used in biodiversity. The following list represents a comprehensive set of objectives for MorphBank:

- Use the full range of capability of MorphBank version 2.5 to include collections and annotations.
- Show the viability of mass upload of data.
- Obtain formal feedback on the Annotation Trials for publication.
- Stress and performance testing.
- Test the security model.
- Test the functionality of the Annotation Model.
- Obtain feedback on the use of the ITIS taxonomic name server.
- Show proof to the National Science Foundation of the progress of the grant.
- Show the viability of using semantically rich annotations.

6.2 Trial Procedures

The trial procedures were quite simple. Lead Scientists and Group coordinators charged with making a presentation of the MorphBank System Version 2.5 to a group of users at an established meeting in advance of the trials. Next, MorphBank system administrators establish user accounts and groups of the participants. MorphBank administrators assisted the scientists in uploading images and related data into MorphBank for which they requested review. A small training session on Determination Annotations was conducted to familiarize the participants with the software. Finally, the group coordinator released the collections and notified the participants that the trials were to begin and monitored the results within the database.

Table 6-1: MorphBank Contributors

<u>Name</u>	<u>Affiliation</u>
Andy Boring	University of Kentucky
Matt Buffington	SEL/USDA NMNH
Andrew Deans	School of Computational Sciences, FSU
Felix Fontal-Cazalla	Museo Nacional de Ciencias, Naturales, Spain
David Houle	Department of Biological Sciences, FSU
Gail Kampmeier	Illinois Natural History Survey, University of Illinois
Johan Liljebblad	University of California, Riverside
Austin Mast	Department of Biological Sciences, FSU
Jose Luis Nieves-Aldrey	Museo Nacional de Ciencias Naturales (CSIC)
Alan Prather	Michigan State University
Albert Prieto-Marquez	Department of Biological Sciences, FSU
Juli Pujade-Villar	Universidad de Barcelona, Spain
Amanda Roe	University of Minnesota
Fredrik Ronquist	Florida State University – SCS
Palmira Ros-Farrél	Universidad de Barcelona, Spain
Susanne Schulmeister	American Museum of Natural History
Michael Sharkey	University of Kentucky
Lars Vilhelmsen	Zoological Museum - Entomology Department
Martin Wiemers	Department of Population Ecology, Faculty of Life Sciences, University of Vienna

6.3 Trial Participants

As of the writing of this document, the individuals listed in table 6-1 are registered users within the system and have actively participated in either the deposit of information and/or annotations. This information was obtained by scanning the MorphBank production database and extracting the individuals who are not exclusively MorphBank Administrator and who have submitted specimens, images, or annotations.

6.4 Initial Trial Feedback

The feedback from the annotation trials has been very positive. Numerous other scientists have used a *beta* version of MorphBank and have provided additional comments which have resulted in the current configuration of the software. Upon examination of the resulting annotations, there were instances where the same type of comments appeared routinely especially during a mass annotations. Numerous comments began with “*This identification*” followed by a set of short remarks. Additionally, there were also many determination annotation comments that start with “*var*” or “*Variation*” indicating a variation of the species. Other comments that appeared more than once include “can't be determined to variety given the images available“ or “can't be determined to variety“. A simple learning algorithm would allow the system to remember the association of these type of comments with the user and permit them to re-use the phrases in additional comments. These comments could also trigger the system to generate a Determination Annotation Sub-Type. Grammar, punctuation, spelling and proper capitalization were problematic in all annotations. It is unknown the affect this will have on search techniques on Annotation data.

The trials of MorphBank will continue for the foreseeable future with additional functionality included within the system. There were numerous other observations on the annotation data. For instance, no user entered any XML data into the system or imported XML data or external references during mass upload operations as originally expected. Some legacy XML data was entered. This is still a highly requested capability and users are expected to start taking advantage of this future. FSU Herbarium Legacy data has not been entered into the system. The original plan was to make these legacy annotations and use the XML upload feature to ensure the data could be stored and retrieved accurately. Although this functionality has been tested and validated, no user has taken advantage of it yet. There has been a large number of new taxonomic names

entered into the local copy of MorphBank but supporting documents and requests to the United States Department of Agriculture to update their version of ITIS has not been accomplished at this time but are planned.

6.5 Image and Specimen Data Summary

Currently there are approximately 102,000 Specimen and Image records in MorphBank contributed by the scientists in Table 6-1. Of those, over 60,000 are image records and over 44,000 are Specimens. Since each Image record must have at least one corresponding Specimen record it can be assumed that the number of Specimen records would never exceed that of image. However, our original analysis of the data presumed that scientists would include several images of each Specimen. Although there are several instances throughout the database where a Specimen does have multiple images, most of the records (76%) have only one image. This ratio is expected to increase as the use of MorphBank increases and more scientists add their collection to the system. In this section, the diversity and relationships of the data and the importance to the trials and briefly described.

Of the Specimens entered, there are 878 different species from both the Plant and Animal Kingdom with 554 (63%) of those that required temporary USDA ITIS names to be entered. This was much higher than originally expected by the analysis of the use of the system but was as a result of including the entire herbarium collection database from the Department of Biological Sciences Department at FSU. However, the newer taxonomic determinations were more critical to the use of the Determination Annotation functions for collaboration of identification of the Specimens and could result in a faster acceptance of newer taxon names into ITIS.




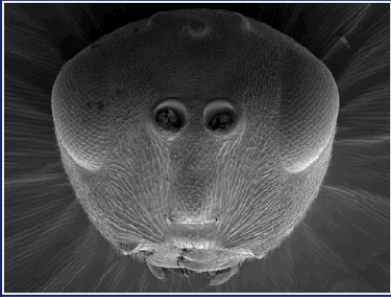





Specimen Record: [67796] <i>Aulacidea tragopogonis</i> 	
<p> Contributor: Johan Lijjeblad </p> <p> Submitter: Johan Lijjeblad </p> <p> Date Submitted: 06-07-2001</p> <p> Last Modified: 06-07-2001</p> <p> Publish Date: 06-07-2001</p> <hr/> <p> Specimen Id: 67796</p> <p> Basis of Record: Specimen</p> <p> Sex: Female</p> <p> Form: Indeterminate</p> <p> Stage: Adult</p>	
<p>Collection</p> <p> Collector: Johan Lijjeblad</p> <p> Institution:</p> <p> Collection Code:</p> <p> Catalog:</p> <p> Collection Num:</p> <p> Date Collected: 0000-00-00 00:00:00</p>	<p>Locality</p> <p> Locality Id: [67795]</p> <p> Locality:</p> <p> Continent: EUROPE</p> <p> Country: UNKNOWN</p> <p> Latitude:</p> <p> Longitude:</p> <p> Precision:</p> <p> Elevation (m):</p> <p> Depth (m):</p>
<p>Determination</p> <p> Class: Insecta </p> <p> Order: Hymenoptera </p> <p> Family: Cynipidae </p> <p> Genus: Aulacidea </p> <p> Species: Aulacidea tragopogonis </p>	<p>Determination Annotations</p>
<p>External links</p>	<p>Other Annotations</p>

Figure 6-1: Specimen Show Record Example

The range of records in the Specimen table included representatives from *Drosophila melanogaster* (Fruit Fly), Hymenoptera (Ants, Wasps, Bees, Saw Flies), Lepidoptera (butterflies and moths), as well as a collection of paleontology data from an FSU hadrosaur (duck bill dinosaur) collection. An example Specimen show record can be seen in Figure 6-1. The number and diversity of the information contained in MorphBank is more than sufficient to validate the utility of the Annotation and other functions, it still represents a fraction of the potential of the system. With over 100,000 Specimen and Image records, the performance of the site has proven to be more than acceptable and should scale easily up to 1,000,000 records.

Collection Record: [104197]
mouthparts sharkey

Submitted by: Michael Sharkey ✉
Submitted date: 08-15-2006
Published date: 02-15-2007
Description:

Show: 20 hits per page Page: Go

◀ 1 2 3 4 ▶ of 4 (80 images)




Image [101923] Archaeoteleia	
View: Head, mouthparts/Posterior Specimen: Female/Adult/Indeterminate Technique: SEM/HMDS, gold-palladium coated	Dim: 1424x1068 [jpg] [tif] Original: TIFF
Image [101922] Archaeoteleia	
View: Head/Posterior Specimen: Female/Adult/Indeterminate Technique: SEM/HMDS, gold-palladium coated	Dim: 1424x1068 [jpg] [tif] Original: TIFF
Image [101884] Belyta	
View: Head, mouthparts/Posterior Specimen: Female/Adult/Indeterminate Technique: SEM/HMDS, gold-palladium coated	Dim: 1424x1068 [jpg] [tif] Original: TIFF

Figure 6-2: Collection Show Example

6.6 Collection Data Summary

Currently, there are 355 Collection objects that reference 15, 240 images. The popularity and immediate use of the Collection feature within MorphBank surprised everyone even when restricting the initial membership in collections to images only. Recall that Collections are designed to hold any MorphBank object but in Version 2.5 membership is restricted to Images for the purpose of validation of the functionality. Figure 6-2 shows an example of Hymenoptera mouth parts created by Dr. Michael Sharkey from the University of Kentucky. An earlier use of the Collection functionality was used by Dr. Fredrik Ronquist and Dr. Johan Liljebblad in exposing the collection of images within MorphBank that were used in a study of phylogenetic characteristics in Hymenoptera.

What can be observed about the makeup of the series of collections created by the different users is the consistency of the relationship of the images to the taxonomic determination of the images. There are not restrictions within MorphBank on the makeup of a collection meaning that any image within system that a user has access to (published or unpublished) is eligible for membership within a collection. It was believed

early in the requirements analysis that scientists would have a diversity of images and data in a collection to display the relationships of the different objects and organisms. However, this feature has not been fully employed at this time. Scientists are using the Collection feature to display unique features of their collection and to share this information remotely with other biologists.

The Collection function is currently being used in the Herbarium Annotation trials. Groups of images are placed in collections and exposed to members of a group to exam and to agree and disagree on their determination. Recall that the only means of performing a mass annotation of multiple images is to place the images within a collection. Dr. Mast has created 19 collections with 1,139 images at the writing of this document composed of images from the Herbarium Collection located at Florida State University. These collections were created for the expressed purpose validating the determination of their taxon. Therein lays the reason for the high relationship of similar genus and species of the images within each collection. More importantly, there exist a sufficient number of collections needed to validate the original requirements for the collection functionality.

6.7 Annotation Data Summary

MorphBank currently contains 148 annotations on Image records from 7 registered research biologists although there are other test records within the database from both MorphBank Administrators and biologists. Annotations were accomplished on images from specimens in both the Animal and plant kingdoms. The following are samples of comments contained in the current set of annotations. What is interesting to note is how often the same type of comments showed up from different users indicating that common phrases and languages are used and could form the basis of a set of semantic annotations. In particular, it was interesting to note how often the mass annotation feature was used on annotations. Additionally it was noted that grammar, punctuation, spelling and proper capitalization were also a problem when allowing free text entry into annotations. It is unknown the affect this will have on search techniques on Annotation data.

- “The sessile, cordate leaf base suggests *A. amplexicaulis* Sm., however the leaf shape and venation are unusual. This is perhaps attributable the

underdeveloped state of the specimen. No other *Asclepias* from the region is a better match”

- “this is the orbicula“
- “The ungues are simple, without subapical teeth. Setae cover the proximal portion only. “
- “Femur is swollen medially. No hairs are present. “
- “This identification may be correct; however, the specimen is depauperate and there is consequently not much to go on. “
- “image inadequate to determine variety only var. *muehlenbergii* is supposed to show up in FL, but without seeing the abaxial face of the perigynium, I can't be sure. “
- “var. *muehlenbergii* “
- “CONFUSION: this is var. *muehlenbergii*, but I can't seem to make the system recognize it as such... editing is a chore“
- “can't be determined to variety“
- “THIS IS VAR. MUEHLENBERGII... previously-indicated problem with annotation“
- “can't be determined to variety given the images available“
- “Note the gynecandrous terminal spike on this individual, atypical in this species. “
- “I made no attempt to distinguish to variety (based on FNA) “
- “Perigynia in some of the pistillate spikes seem narrow for this species“
- “In some of the detail images, the perigynia seem to be rather strongly veined for this species; however, the bulk of images point to *Carex glaucescens*“
- “Annotation related to the following images: 91391, 91392, 91393, 91395, 91396, 91397, 91398, 91399 of Collection id [109108] “

- “This identification may be correct; however, the specimen is depauperate and there is consequently not much to go on. “
- “This identification may be correct; however, the specimen is depauperate and there is consequently not much to go on. Identification may be correct, but specimen is depauperate. “
- “The sessile, cordate leaf base suggests *A. amplexicaulis* Sm., however the leaf shape and venation are unusual. This is perhaps attributable the underdeveloped state of the specimen. No other *Asclepias* from the region is a better match. “
- “this is the orbicular“
- “The ungues are simple, without subapical teeth. Setae cover the proximal portion only. “
- “Femur is swollen medially. No hairs are present. “

There were numerous other observations on the annotation data worth noting. In particular no user entered any XML data into the system even though this was identified early in the requirements analysis as a necessity in making MorphBank compatible with other systems. There are also very view general or legacy comments despite the fact that these were features specifically requested by some of the biologists involved with the project. A great deal of effort was placed into allow users to specify a variation of the taxon determination by using the formal prefix and suffix latin names used by biologists. However, these too were not used. This is probably due to the fact that researchers involved with the trials were given an abridged version of the MorphBank user’s manual in the form of a step by step method to enter annotations that deliberately avoid the other features.

Annotation Record: [109845] Title = femur

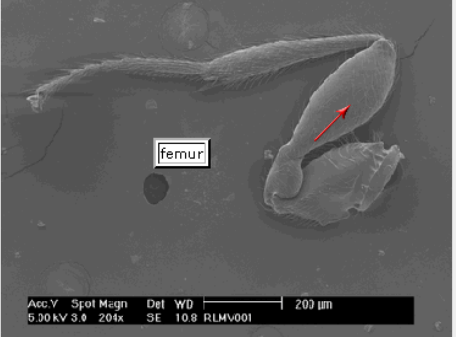
<p>Contributed By: Andrew Deans ✉</p> <p>Date Contributed: 10-07-2006</p> <p>Last Modified: 10-07-2006</p> <p>Publish Date: 04-06-2007</p> <hr/> <p>Specimen Id: [78506]</p> <p>Sex: Male</p> <p>Collector: B. L. Fisher et al.</p> <p>Species Name: Ceraphron sp.RLMV005</p> <p>Object Id: [78554]</p> <p>Object Type: Image</p> <p>Type of Annotation: General</p>	
<p>Comments</p> <p>Femur is swollen medially. No hairs are present.</p>	<p>Related Annotations to this image</p> <p>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]</p> <p>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]</p>
<p>General Data</p>	

Figure 6-3: Sample Single Image Annotation

Feedback from users on the Annotations came to this research in the form of personal comments or on the MorphBank twiki site located at the following URL with restricted access. (<https://www.scs.fsu.edu/twiki/bin/viewauth/MorphBank/FeedBack>). The following comments are from one of the MorphBank research Scientists, Dr. Andrew Deans (see Figure 6-3 for an example single image Annotation).

- “the current setup for the annotation arrow placement makes it really difficult to easy point to features in the middle of the frame. For example: Image Record [100628] has a large 'bump' between the two claws. I would like to point to it, but the arrows always end up pointing in the opposite direction. Is it possible to select which end to place the arrow on the line, like you can in PowerPoint? Example: <http://morphbank.net/Show/?id=109843>”

- “I almost think that the arrow should ALWAYS appear on the opposite to where it appears now. I also had difficulty with this annotation in terms of placing the arrow without obstructing other features: <http://morphbank.net/Show/?id=109845> “
- “if I submit a screwy annotation and MB tells me it's no good (some error about how I didn't provide the right info) I have to do the annotation over again - add an arrow, select the kind of annotation, etc. Could there be some memory here, so all I have to do is correct the one thing I overlooked? “
- “I chose to use a yellow arrow for these annotations, but it's the red arrow that shows when I view them: <http://morphbank.net/Show/?id=109844>, <http://morphbank.net/Show/?id=109845>”

6.8 Determination Annotation Data Summary

A Determination Annotation is a special type of annotation that includes additional data concerning the determination of a specimen. A Determination Annotation extends the Annotation class therefore the number of Determination Annotations cannot exceed those of Annotations. During the course of the trials, the MorphBank system has amassed 136 determination annotations out of a total of 148 Annotations. More than 91% of all annotations were of this type. The rest were general annotations, most of which were test records.

Of the 136 Determination Annotations, 11 were test records. All of the remaining records were in agreement with the original determination which was that assigned to the Specimen Record. Recall in a previous section that the Specimen record was considered to be a determination annotation for the sake of statistics in the Annotation Module. There were no disagreements which was surprising considering the number of entries. Additionally, there were a relatively few (4 records) that included a prefix or suffix other than “none”. Figure 6-4 shows a sample determination annotation from the test collection illustrating the flexibility and capability of the data gathering tool.

Annotation Record: [109156] Title = Determination

Contributed By: Andrew Hipp 
Date Contributed: 09-08-2006
Last Modified: 09-08-2006
Publish Date: 10-16-2006

Specimen Id: [84663]
Sex: Undetermined
Collector: R. Kral
Species Name: Carex glaucescens
Object Id: [90501]
Object Type: Image
Type of Annotation: Determination



Comments

In some of the detail images, the perigynia seem to be rather strongly veined for this species; however, the bulk of images point to Carex glaucescens

Related Annotations to this image

TITLE: Determination
TYPE ANNOTATION: Determination
BY: Andrew Hipp
DATE CREATED: 2006-09-08 16:35:25
RELATED ANNOTATIONS OF ID: [109156]
SINGLE SHOW OF ANNOTATION ID: [109156]

Related Annotations1

Determination Data

Specimen Id: [84663]
Taxonomic Serial Number: [39396]
Taxonomic Name: [Carex glaucescens]
Prefix: [none]
Suffix: [none]
Type Determination: [agree]
Source of Id: [Andrew Hipp]
Resources used in Id: [Flora of North America]
Materials used in Id: [Image]

Taxonomic Name	Taxon Author	Prefix	Suffix	A	D	S
Carex	Ell.	none	none	2	0	1

Figure 6-4: Sample Determination Annotation

The uses of the mass determination annotations were the vast majority of 98 records from 12 different collections. This indicates that the grouping of images and specimens within like taxonomic areas has proven to be a very useful tool in this early

version of MorphBank. Figure 6-5 shows an example of a mass annotation. What is interesting to note about the difference between two different type of annotations is that MorphBank is able to display the difference set of data. The comments section in Figure 6-5 are automatically populated with the identification numbers of all the related images in the Collection with the determination Annotation. Also, in all Determination Annotations, the last area of the screen displays all related annotations to this particular image/specimen. A few users were able to efficiently and affectively generate a large amount of reliable data that can easily be searched and discovered by other MorphBank users.

6.9 Chapter Summary

This chapter discussed the processes by which the research behind the design and creation of a phylogenetic image database was accomplished for the purpose of demonstrating the feasibility of using a complex annotation tool to easily add complex semantically related information to existing data. At the writing of this dissertation, MorphBank version 2.5 is still being used by research scientists around the world and additional annotation trials are planned. Additionally, the MorphBank team is planning to gather survey data from the Annotation trials in an attempt to improve the human-computer interface and increase the reliability and accuracy of the MorphBank objects. Additional research is also being conducted independently of the Annotation trials to retrieve data from the actual source of the information. There are large repositories that exist that already store related MorphBank objects such as complex specimen data, locality information, and publications that can be retrieved at their source rather than replicated in MorphBank thereby increasing reliability and accuracy while decreasing storage requirements.

Annotation Record: [109770] Title = Determination

Contributed By: Mark Fishbein 
Date Contributed: 09-24-2006
Last Modified: 09-24-2006
Publish Date: 10-16-2006

Specimen Id: [91652]
Sex: Undetermined
Collector: Steve L. Orzell, Edwin L. Bridges
Species Name: *Asclepias humistrata*
Object Id: [92093]
Object Type: Image
Type of Annotation: Determination



Comments

Annotation related to the following images: 92091, 92092, 92093, 92094, 92095, 92096, 92097, 92098, 92099, 92100 of Collection id [109274]

Related Annotations to this image

TITLE: Determination
TYPE ANNOTATION: Determination
BY: Mark Fishbein
DATE CREATED: 2006-09-24 20:18:13
RELATED ANNOTATIONS OF ID: [109770]
SINGLE SHOW OF ANNOTATION ID: [109770]

Related Annotations 1

Determination Data

Specimen Id: [91652]
Taxonomic Serial Number: [30272]
Taxonomic Name: [*Asclepias humistrata*]
Prefix: [none]
Suffix: [none]
Type Determination: [agree]
Source of Id: [Mark Fishbein]
Resources used in Id: [My Expert Opinion]
Materials used in Id: [Image]

Taxonomic Name	Taxon Author	Prefix	Suffix	A	D	S
<i>Asclepias</i>	Walt.	none	none	2	0	1

Figure 6-5: Sample Mass Determination Annotation

CHAPTER 7

CONCLUSION

The problem of using a biodiversity information system for storing semantically rich information in a highly collaborative environment was solved during this research.. Large amounts of data including images were cataloged and stored in a well organized and well designed web-based information system and made available to a large group of research scientists for their use and feedback. A complex and standardized schema was designed and implemented that permits scientists to store related objects, their association, and the context together in an easily searchable format. Finally, an environment was developed for scientists to annotate their specimens and remotely collaborate their research with their colleges in a fashion never before seen in the biodiversity community. The unique methods and ideas developed in this research have made a significant contribution to the computer science and biodiversity systems. Separate organizations and groups of individuals were given their own research space and the ability to modify a range pf definitions within MorphBank to suite their own personal research goals. This research has shown that by developing a secure environment for such collaborations, scientists have greater confidence in the results. The results of queries are returned significantly faster, more reliable, and more in context than a similar “google-type” plain text search. Complex internal and user-defined relationships were allowed to form within the system. Finally, the research has shown that the resulting methods provide a good utility that has significantly improved the current methods and have added insight into the needs and desires for biodiversity systems.

The current MorphBank project is now taking lessons learned and applying this towards future versions of the software. The system has been presented at numerous conferences, workshops, working groups, and meetings. Additionally new users, groups, images, and annotations are being added every day. Plans exist to propose to the biodiversity community a MorphBank Consortium of interested scientists and organizations that will continue manage the future direction of the project. Other projects have already started placing external MorphBank object references in their own databases. A search on the internet has revealed numerous links that are using the

MorphBank Show function and Collection function to display their works and using the external link function to associated MorphBank objects with external data libraries.

7.1 Research Results

The goal of this research was to develop and implement methods for scientists to search annotated databases for information they need to support their work and to share work, collaborate with their colleges, and share their research with the scientific community. This was not only accomplished but the resulting MorphBank version 2.5 system is now a production quality software product used by hundreds around the world. Separate organizations with different methods and architectures can now import their data into MorphBank, reference their own databanks and collaborate their findings with their colleges. With the implementation of the Darwin Core standard, scientists can determine exactly what data is needed. MorphBank abstracts the location of the data and knowledge of the physical location by the user of the data is no longer required. The user interface handles the extraction of the appropriate data. Transformations are accomplished automatically through the MorphBank Show function.

The idea of ontology was clearly defined in MorphBank through the complex associations of triplets within the system. Objects are related to other objects within MorphBank through a relationship and a context. When searching in the database for wings associated with all related objects for insects, the system understands that only “wings” associated with insects are to be searched. This concept is extended in annotations where the user can find all associated objects in the same context of collections, taxonomic names, species, and related annotations.

7.2 Research Objectives Achieved

The approach to solving the problem of annotations in biodiversity information systems was accomplished using several features. The complex schema structure using the Darwin Core standard and the Integrated Taxonomic Information System (ITIS) solved the problem of identifying and properly naming the traits in the database system so they could be easily identified. The concept of the Collection solved the next problem of relationships. Since all MorphBank objects are centrally cataloged, detailed and complex relationships can be built that are unique to each individual scientist. Annotations on these collections can tie the objects together in a common context where

the related collections can be found when searching for each individual object. Feedback from the initial users of these features have been extremely favorable. The combination of a biodiversity image system combined with a highly collaborative environment with determination annotations has been successful.

7.3 Annotation of Metadata Relationships

In most scientific disciplines, the gathering and analysis of data represents a significant portion of the activities of the scientists involved with the research. The data is then stored, analyzed, transformed, cataloged, and annotated in a variety of formats ranging from fixed length flat files to relational databases. Data objects have relationships that must be identified and annotated in order to provide validity to the scientific research. Informal data repositories tend to be unorganized into hand written ledgers or log books and contain no patterns needed for quick and accurate searches. Recording, understanding, and retrieving information contained in most scientific annotation is one of the major challenges facing the science today.

The semantic association annotation tool in MorphBank takes a more logical approach to solving this problem by making the data available to the world via a view point that would allow for discovering and retrieval using simple methodologies. Database tables that are divided (normalized) into separate tables for the sake of organization sometimes lose meaning because of functional decomposition. The schema design incorporates several features that allow the association of objects within the information system given a context and a relationship that helps the discovery of metadata.

7.4 Verification of Research

As mentioned earlier in this document, the requirements for the MorphBank system were difficult to document and validate because of the diverse opinions among the various experts in the field on such subjects such as required data items, common reports, taxonomic name servers, etc. By analyzing the actual use of the system itself, the needs and desires of the users within the biodiversity community can be determined with a great deal certainty. This section briefly reviews the information presented to this point to show the magnitude of the success of the research and to suggest features for future versions of MorphBank.

7.4.1 User and Group Relationships

MorphBank currently has 106 registered users spread out in 95 groups. Individuals conducting annotation trials are creating separate groups for review of the collection of images maintaining ownership and permitting other scientists in the group to view the collections, create new collections of their own under the ownership of the group, and annotate the objects. The “Plants” group has a total of 21,948 different objects that they own, most of which are from the herbarium collection at FSU. It is anticipated that other research departments will follow suite and declare ownership of their collection under a single group and take full advantage of the functionality of the security module and object sharing features built into MorphBank.

7.4.2 Annotation and Related Determination Annotations

The Annotation trials to date have shown that an integrated semantic annotation tool is not only a viable component to a scientific data but one that adds a great deal of utility by replacing the need of scientists to physically collaborate on research projects and associated materials. This research found the definition of *annotation* changed as detailed aspects of the data collection practices of the biodiversity systems were uncovered. There were several questions concerning concepts in annotations such as:

What is a legacy annotation? Is it an annotation made before MorphBank was conceived or was it an annotation made outside of MorphBank but added at a later date? Annotations are always conceived (however briefly) before they are entered into MorphBank. Technically, all annotations could be considered legacy. We therefore define legacy annotations as those annotations entered previously into another system (either hardcopy or electronic).

How is legacy data entered into MorphBank that does not conform to Darwin Core standards? Dates from older specimens may not have specific days but may be something along the line of “Fall 1896” or “June 1922”. Additionally, the actual collection date of a specimen may not be known at all. Some insect traps are left for weeks before they are collected. The exact date the insect fell into the trap is not know. Mapping of legacy data into the MorphBank schema can be accomplished through legacy annotations and the use of customized XML Schemas and partial documents.

Should older but out of date taxonomic names be used to preserve the historical aspect of the legacy annotation? Names coined 100-200 years ago may no longer be applicable or specimens may be properly identified. Should these names be corrected or should the data from a historical record be shown? Again, we use legacy determination annotations to store historical data such as this.

Should mass determination annotations only affect the specimen objects? Right now as of version 2.5, images are annotated and when a determination annotation the specimen id is placed in the Determination Annotation record but the annotation is still associated with the image. Should MorphBank develop a strategy where users self define annotations and define what objects should be related? This capability was placed in the initial design and can be implemented with little effort.

The original design of the semantic association annotation tool placed great emphasis on the physical markup of images. However, as the research progressed it became apparent the relationship of the annotated object (image) with other types of data was significantly more important than originally realized and more important than the ability to place markers on the image. This is seen from that fact that of the 148+ annotations entered into the system only three employed this feature. It is believed that phylogenetic character/state annotations will have more need of this feature.

Finally, the combination of collections and mass annotations was a late realization in the MorphBank project but one that has opened many doors to other types of annotations and relationships. This feature could prove to be truly useful in automatically defining relationships in objects by following the object-to-object relationships along the same context. Various different specimens can be associated with each other by obtaining the membership of images and specimens in collections with the intersection of the membership within those collections. These objects can also be followed to their respective localities, collectors, publication, and taxonomic determinations.

7.4.3 Reliability and Rating of Annotations

All of the scientists involved with the Annotation trials were rated as Lead Scientists and are given Privilege Taxonomic permissions at very high levels. Additionally, the person who owns the data within the system are also highly rated and thus the reliability of the specimen and images are going to be equally high. The

determination annotations made on the herbarium collection series were all positive identifications and the consistency of these ratings also cause the reliability and the ranking of the data to consistently ranked. The difference caused in the ratings of the different specimens comes into play because users assigned to the Plants group were given a Privilege Taxonomic rating at the Kingdom (Plant -240) level while other groups involved with insects were give a rating at the Order level (Hymenoptera – 100). While the consistency of the ratings among the groups remains constant, comparing the reliability between different groups or even of the entire system is not.

What can be concluded from an analysis of the reliability data is that the quality of the submissions to date and the qualifications of the individuals responsible for the data are of a high caliber. This is accomplished by running a simple script that uses the ratings of the user (U), plus their taxonomic privilege (T), plus their primary qualification (Q), plus the quality of the specimen or data (D). What can be see is that members of the MorphBank development team have a low user rating and as we would expect the data inserted by them is of a low value. To prove this assertion we need only examine all of the data within MorphBank to assess the quality of the information and we find a direct correlation between low quality information and that data entered by the MorphBank development team for testing.

7.4.4 Semantic Relationship of Annotations

One of the more radical features examined during the course of this research is the feasibility of a semantic annotation capability that allows users of MorphBank additional capability to store, find, and retrieve information. The process involved marrying the power of a relational database with the flexibility of an XML document.[GZMRR06] Simply what was accomplished was that MorphBank users have the option in the current version to store XML documents with annotations. Future versions may allow XML documents to be stored with any object. Along with the XML document, the associated schema can be stored that describes the content of the XML document. The XML Schema is stored as an Annotation Type. The problem: When searching for data in comments on annotations the only method that is currently available is free text search of ALL comments. Search a vernacular names of “fruit” and “fly” would result in all records that had any occurrence of the two words. Comments

concerning plants or animals that have a “fly” as a pest and eat “fruit” will result in positive match. Figure 7-1 shows an example of how a search for such a string could result in an undesirable result. What is desired is a more exact match to the specific meaning of the words that results in fewer records that are closer to what was desired.

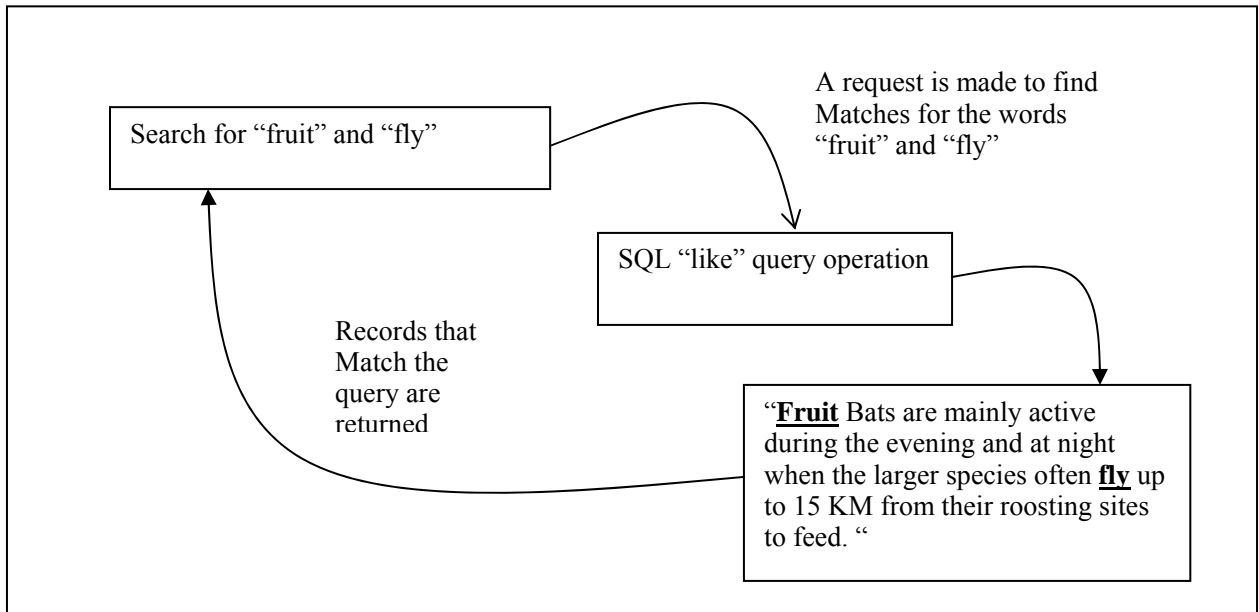


Figure 7-1: Simple Text Search

With test XML documents in MorphBank already, a simple test search feature was written accept a semantic query. This query includes both the attribute and the desired text search. Figure 7-2 displays an example of a search for a specific XML attribute, comparison operator, and search string. The experimental software developed as proof of concept in MorphBank version 2.5 searches the type annotations to determine which XML Schemas contain the attribute “Vernacular”. The database then extracts only those records that contain those schemas and the attribute <Vernacular> is search for the appropriate string and operation (=, <=, >=, <, >, or !=). As found in the benchmark test, only a fraction (approximately 5%) of the annotation text was actually search versus performing a straight text search on all annotation comments. The search is being pared down by (1) reducing the number of entities that are searched, and (2) by searching on the text identified as being part of the attribute requested. Accuracy is much higher then

in a standard MorphBank text search because extraneous matches are not revealed. This varies depending upon the amount of XML data that is included with annotations.



Figure 7-2: XML Search Screen

Figure 7-3 shows the results of the XML search function by displaying, in this case, the annotations that satisfy the results. The records id (110238) allows the user to bring up a window of the actual annotation while the Object ID (63952) is a hot link to the actual object being annotated. In this case an image.

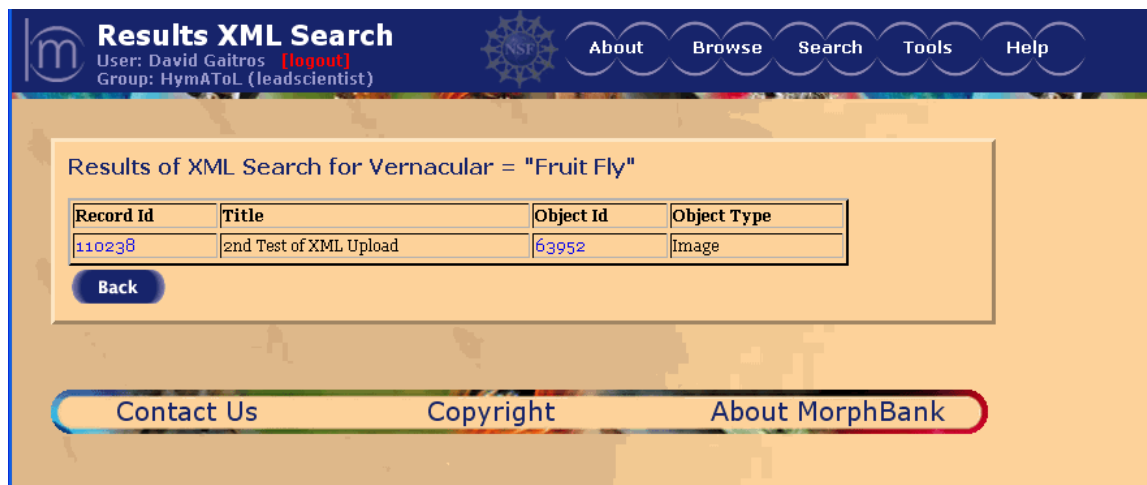


Figure 7-3: XML Search Results Screen

7.5 Accomplishment of Research Challenges

There were several major hurdles that had to be overcome during the course of this research. One of the early challenges faced by the MorphBank Research team was the issue of different ontologies and research methodologies among the different communities. Although this problem was not completely solved the research was able to discover a significant amount of common areas and agree on the format of a communication dialogue in the form of the ABCD standard and XML documents. This actually proved to be better than had been originally expected.

Another issue was the transformation of annotation data from a manual method to an automated/digital form. There are automated annotations and object relationships that could be created without user intervention such as capturing the metadata from image tags. By making the underlying annotation model flexible, a forum can be created where a vast amount of diverse data could be entered into MorphBank and reasonably searched. Each research discipline could define (through an XML schema) their different ontologies and thus provide a means to express their ideas and thoughts to each other. Researchers can thus learn the meaning of other ontologies by their association with known objects in a database.

The single major challenge of the research project was the magnitude of work that had to be accomplished to prove the results. Despite the massive amount of software that had to be designed, written, and tested; there was an equally important task of importing large amounts of high quality biological data from existing collections into MorphBank order to gather sufficient information to formulate a conclusion, a trial of the prototype software for use in the annotation trials. Thanks to a very generous NSF BDI research grant and a dedicated development team, Version 2.5 was released July 29th, 2006 in time for use in a remote annotation trial for a herbarium collection organized by Dr. Austin Mast. Another annotation trial of hymenoptera specimens will be conducted by Dr. Fredrik Ronquist in the near future. The annotation software is connected to a fully functional biodiversity system currently under maintenance and configuration control.

7.6 Future Work

There are several important initiatives that are being considered for the future of the project. Some are related to Annotations and others are related to migrating MorphBank to a more open architecture. The following are some of the suggested areas:

- Continue to extend the capability of Annotations and Collections
- Turn on the feature that allows for the annotation of any object
- Turn on the feature that allows for any object to be in a collection
- Research more efficient search techniques for semantic associations
- Complete development and release of phylogenetic character state software
- Research the possibility of further developing the extensible schema capability
- Analysis of the complexity of relationships of the objects associated through collections and annotations
- Expand and mature the use of Life Science Identifiers
- Implement a security strategy that is separate from the implementation of the software
- Map the current data schema to the ABCD standard for the purpose of exporting data.
- Publish results in high quality journal. Continued exposure at conferences and workshops

APPENDIX A MORPHBANK CONTRIBUTORS

The following individuals were responsible, at least in part, for the research and development behind the MorphBank Project.

Principle Investigators	Department
Dr. Fredrik Ronquist	School of Computational Science
Dr. Greg Riccardi	College of Information
Dr. Austin Mast	Department of Biological Sciences
Dr. Corrine Jorgensen	College of Information
Dr. Peter Jorgensen	College of Information
Dr. Robert van Engelen	Department of Computer Science
Dr. Greg Erickson	Department of Biological Sciences

Developers	Role/Responsibilities
David A. Gaitros	Project Manager, Software Design, Annotations
Wilfredo Blanco	Graphics, Analysis and Design
Neelima Jammigumpula	Sys Admin, Database Design
Karolina Maneva-Jakimoska	Java Programmer, Analysis and Design
Steve Winner	Web and Database Design
Cynthia Gaitros	Publishing and Technical Writer
Debbie Paul	Biology Functional Analyst
Katja Seltmann	Biology Functional Analyst

Research Associates	Department
Dr. Gordon Erlebacher	School of Computational Sciences
Dr. Andrew Deans	University of Illinois
Dr. Matthew Buffington	Systematic Entomology Laboratory`
Shayne Steele	Annotation Technology

Student Research Asst.	Contribution
Gabriel Logan	Display related image annotations
Jason Simmons	Display image annotation
Stanislov Ustymenko	Web Services
Wei Zhang	Security Web Services
Alison von Eberstien	MorphBank Requirements
Janet Capps	MorphBank Requirements

CEN 4010 Class Project Summer 2004	Contribution
Duane Griffiths, Antoineet S. Mulai, Richard Cook, Yuval Peress, David Kopicki, Demetrius Brown	Web Site Requirements and Initial Design
Thomas Bonfield, Christi Shirley, Keith Zenoz, Michael Jason, Gabriel Logan, Robert Worley	Database Schema and Low Level Software Library Prototype Development
Christopher Albritton, Nikki Brown, Johan Martinez, Kerstin Galutera, Kowit Jitraphai	Annotation Technology Prototype
Michael Lind, Justin Christofoli, Daniel Beech, Saif Mazhar, Joe Barrett, Jesse Levier	MorphBank Operational Specifications
Masa James, Bruce Bayha	Server, Backup System, and Mirror Site Configuration Recommendation
Erica Bourne, Luisa Pleger	Trademark and Copyright
Theresa Pace	Documentation, Configuration and Management

APPENDIX B DARWIN CORPS STANDARD 2.0

This document is an up-to-date specification of elements and terms used by the MorphBank Database Management system. Some of the terms are Darwin core elements and can be found at <http://gbif.nbi.gov/standards/standards.html> web site.

1. **ID:** An 18 Character field generated by the MorphBank system for each object entered into the database. Partially compliant with the proposed LSID standard, the ID is comprised of a namespace (MorphBank) and a unique 8 digit serial number generated by the MorphBank system. The id is cataloged with the database for object searches.
2. **DateLastModified:** ISO 8601 compliant stamp indicating the date and time in UTC(GMT) when the record was last modified. Example: the instant "November 5, 1994, 8:15:30 am, US Eastern Standard Time" would be represented as "1994-11-05T13:15:30Z"
3. **InstitutionCode:** A "standard" code identifier that identifies the institution to which the collection belongs. No global registry exists for assigning institutional codes. Use the code that is "standard" in your discipline.
4. **CollectionCode:** A unique alphanumeric value which identifies the collection within the institution.
5. **CatalogNumber:** A unique alphanumeric value which identifies an individual record within the collection. It is recommended that this value provides a key by which the actual specimen can be identified. If the specimen has several items such as various types of preparation, this value should identify the individual component of the specimen.
6. **ScientificName:** The full name of lowest level taxon the cataloged item can be identified as a member of. This includes genus name, specific epithet, and subspecific epithet (zool.) or infraspecific rank abbreviation, and infraspecific epithet

(bot.). Use the name of suprageneric taxon (e.g., family name) if the cataloged item cannot be identified to genus, species, or infraspecific taxon.

7. BasisOfRecord: An abbreviation indicating whether the record represents an observation (O), living organism (L), specimen (S), germplasm/seed (G), etc.
8. Kingdom: The reference to the kingdom to which the organism belongs.
9. Phylum: The reference to the phylum (or division) to which the organism belongs.
10. Class: The reference to the class name of the organism.
11. Order: The reference to the order name of the organism.
12. Family: Indicates the family name of the organism.
13. Genus: References the genus name of the organism.
14. Species: The reference to the specific epithet of the organism.
15. Subspecies: Indicates the sub-specific epithet of the organism.
16. ScientificNameAuthor: The reference to the author of a scientific name. Author string as applied to the accepted name. It can be associated with more than one author (concatenated string). It should be formatted according to the conventions of the applicable taxonomic discipline.
17. IdentifiedBy: The reference to the name(s) of the person(s) who applied the currently accepted Scientific Name to the cataloged item.
18. YearIdentified: The year portion of the date when the collection item was identified. It is entered as four digits [-9999..9999], e.g., 1906, 2002.

19. MonthIdentified: The month portion of the date when the collection item was identified. It is entered as two digits [01..12].
20. DayIdentified: The day portion of the date when the collection item was identified. It is entered as two digits [01..31].
21. TypeStatus: Indicates the kind of nomenclatural type that a specimen represents. In particular, the type status may not apply to the name listed in the scientific name, i.e. current identification. In rare cases, a single specimen may be the type of more than one name.
22. CollectorNumber: An identifying "number" (really a string) applied to specimens (in some disciplines) at the time of collection. Establishes and links different parts/preparations of a single specimen and between field notes and the specimen.
23. FieldNumber: A "number" (really a string) created at collection time to identify all material that resulted from a collecting event.
24. Collector: The name(s) of the collector(s) responsible for collecting the specimen or taking the observation.
25. YearCollected: The year (expressed as an integer) in which the specimen was collected. The year should be entered as four digits (e.g. 1972 must be expressed as "1972" not "72").
26. MonthCollected: The month of year the specimen was collected from the field. The month should be entered as two digits (Possible values range from 01...12 inclusive).
27. DayCollected: The day of the month the specimen was collected from the field. The day should be entered as two digits (Possible value ranges from 01..31 inclusive).
28. JulianDay: The ordinal day of the year; i.e., the number of days since January 1 of the same year. (January 1 is Julian Day 1.)

29. TimeOfDay: The time of day a specimen was collected expressed as decimal hours from midnight local time (e.g. 12.0 = mid day, 13.5 = 1:30pm)
30. ContinentOcean: References the continent or ocean from which a specimen was collected.
31. Country: Indicates the country or major political unit from which the specimen was collected. ISO 3166-1 values should be used. Full country names are currently in use. A future recommendation is to use ISO3166-1 two letter codes or the full name when searching
32. StateProvince: The state, province or region (i.e. next political region smaller than Country) from which the specimen was collected.
33. County: The county (or shire, or next political region smaller than State/Province) from which the specimen was collected.
34. Locality: Indicates the locality description (place name plus optionally a displacement from the place name) from which the specimen was collected. Where a displacement from a location is provided, it should be in un-projected units of measurement.
35. Longitude: References the longitude of the location from which the specimen was collected. This value should be expressed in decimal degrees with a datum such as WGS-84.
36. Latitude: Indicates the latitude of the location from which the specimen was collected. This value should be expressed in decimal degrees with a datum such as WGS-84.
37. CoordinatePrecision: An estimate of how tightly the collecting locality was specified; expressed as a distance, in meters, that corresponds to a radius around the latitude-longitude coordinates. Use NULL where precision is unknown. This value cannot be estimated, or is not applicable.

38. BoundingBox: This access point provides a mechanism for performing searches using a bounding box. A bounding box element is not typically present in the database, but rather is derived from the latitude and longitude columns by the data provider.
39. MinimumElevation: Indicates the minimum distance in meters above (positive) or below sea level of the collecting locality.
40. MaximumElevation: The reference to the maximum distance in meters above (positive) or below sea level of the collecting locality.
41. MinimumDepth: The minimum distance in meters below the surface of the water at which the collection was made; all material collected was at least this deep. Data is positive below the surface, negative above (e.g. collecting above sea level in tidal areas).
42. MaximumDepth: The maximum distance in meters below the surface of the water at which the collection was made; all material collected was at most this deep. Data is positive below the surface, negative above (e.g. collecting above sea level in tidal areas).
43. Sex: References the sex of a specimen. The domain should be a controlled set of terms (codes) based on community consensus. Proposed values: M=Male; F=Female; H=Hermaphrodite; I=Indeterminate (examined but could not be determined; U=Unknown (not examined); T=Transitional (between sexes; useful for sequential hermaphrodites).
44. PreparationType: Indicates the type of preparation (skin, slide, etc). It is probably best to add this as a record element rather than access point. It should be a list of preparations for a single collection record.
45. IndividualCount: The number of individuals present in the lot or container. This is not an estimate of abundance or density at the collecting locality.

46. PreviousCatalogNumber: The previous (fully qualified) catalog number of the cataloged item if the item earlier identified by another catalog number, either in the current catalog or another institution / catalog. A fully qualified catalog number is preceded by an institution code and collection code, with a space separating the each sub element. Referencing a previous catalog number does not imply that a record for the referenced item is or is not present in the corresponding catalog, or even that the referenced catalog still exists. This access point is intended to provide a way to retrieve this record by previously used identifier, which may be used in the literature. In future versions of this schema this attribute should be set-valued.
47. RelationshipType: A named or coded valued that identifies the kind relationship between this collection item and the referenced collection item. Named values include: "parasite of", "epiphyte on", "progeny of", etc. In future versions of this schema this attribute should be set-valued.
48. RelatedCatalogItem: The fully qualified identifier of a related catalog item (a reference to another specimen); institution code, collection code, and catalog number of the related cataloged items, where a space separates the three subelements.
49. Notes: Free text notes attached to the specimen record.

Additional Resources

- ❖ <http://digir.net/schema/conceptual/darwin/2003/1.0/darwin2.xsd>
- ❖ <http://tsadev.speciesanalyst.net/documentation/ow.asp?DarwinCoreV2>

APPENDIX C MORPHBANK VERSION 2.5 ANNOTATION USERS MANUAL

Annotation

Annotation allows users to add additional information to objects in the MorphBank relational database. An annotation is a comment about an object (usually an image or collection) that is stored separately from the object itself. Annotations are identified in MorphBank by a unique internal id.

The created annotations are published (viewable to the world) when released by the creator (default 6 months if not otherwise notified).

Note: Initially, only images and specimens have annotation options but in future versions, users will be able to annotate any MorphBank object (i.e. image, specimen, locality, view, publication, annotation, etc).

Guidelines for working with annotations:

A user may have multiple annotations that will be identified by a title on the screen. Since the annotation will have a unique internal identifier, the name may be duplicated but is not recommended. (When making mass annotations all will have the same initial title in the annotation manager.)

Any logged in user can annotate any image or collection that is released. Any logged in user can annotate any image or collection that has not been released provided they belong to the group who owns the image or collection.

Unpublished owned annotations:

- A user may edit the makeup of their own unpublished annotations.
- A user may delete an unpublished, owned annotation.

Unpublished annotations owned by other users:

- A user may browse unpublished annotations of other users within groups to which he/she belongs.
- A user may view unpublished annotations of other users within groups to which he/she belongs.

Published annotations:

- Published annotation cannot be edited.
- Published annotations are viewable to the world.

The user's group/user's annotation relationship:

- The user's annotation will be shared with a group in MorphBank. The user must declare which group they belong before they create the annotation (declared through **Select Group** in the login process) and that annotation is shared with the declared group.
- The annotation will be immediately viewable to all users in that group (The annotation cannot be accessed by the world until it is published).
- Although the owner may edit their own unpublished annotation, other members of the group may not.

Annotation Manager

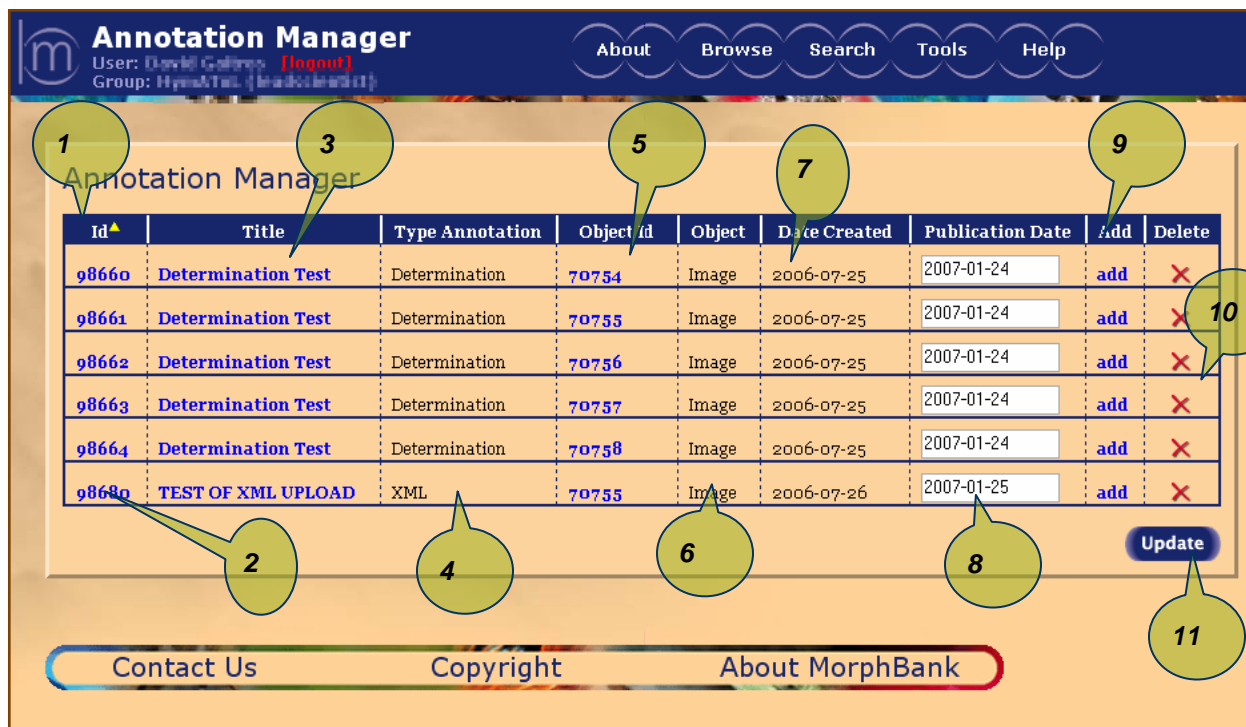
The **Annotation Manager** offers the user a list of all the annotations that have been created under the current username and group (Figure 1). There is no limit on the number of annotations a user may have. To access other annotations under the same username but created under another authorized group, return to the **Select Group** screen and login under that group.

Screen Use Tip:

To return to **groups**, click the **tools** button on the page header and choose **select group** from the list.



The **Annotation Manager** (manager of the user's personal annotations) is directly accessed by choosing **Annotation Manager** from the **Tools** menu (located on the opening MorphBank screen or on the page header).



Id	Title	Type Annotation	Object Id	Object	Date Created	Publication Date	Add	Delete
98660	Determination Test	Determination	70754	Image	2006-07-25	2007-01-24	add	X
98661	Determination Test	Determination	70755	Image	2006-07-25	2007-01-24	add	X
98662	Determination Test	Determination	70756	Image	2006-07-25	2007-01-24	add	X
98663	Determination Test	Determination	70757	Image	2006-07-25	2007-01-24	add	X
98664	Determination Test	Determination	70758	Image	2006-07-25	2007-01-24	add	X
98680	TEST OF XML UPLOAD	XML	70755	Image	2006-07-26	2007-01-25	add	X

Figure 1 Annotation Manager

This Figure Contains Test Data

Tag 1-**Annotation manager header**: Click on the column headers to sort the list of the applicable data by number order, alpha order, date order, etc.

Tag 2 -**Annotation id**: This is a MorphBank issued identifier. Click on it to view the associated annotation.

Tag 3 -**Annotation title**: Clicking on this title will take the user to the **Edit Annotation**

screen (Figure 2). This screen contains the previously entered annotation data that can be edited by the owner. Take note that the type of annotation can not be altered. (**Edit Annotation** is only available to the owner if the annotation is not yet published.) Complete instructions on this area can be found in the [Edit Annotation](#) area of this manual.

Figure 2 Edit Annotation

This Figure Contains Test Data

Tag 4 -**Annotation type**: There are currently four types of annotations possible: **Determination, General, Legacy** and **XML** (see **Types of Annotations** later in this chapter.)


Tag 5 -**Object id**: This represents the identifying number of the object (image, specimen, etc.) being annotated. Clicking on the id will take the user to the **Single Show** screen (Figure 3) that displays the record which contains the image and related information.

Figure 3 Single Show Image Record

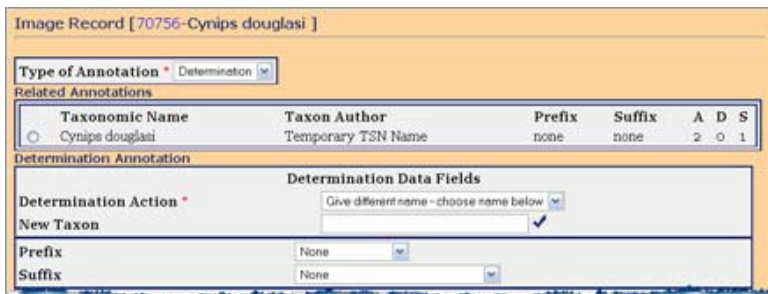
Tag 6 -**Type of object being annotated**: Initially, only images and specimens will have annotation options but in future versions, users will be able to annotate any MorphBank object (i.e. image, specimen, locality, view, publication, annotation, etc).

Tag 7 -**Date created:** This is the date that the annotation was submitted to MorphBank. It is automatically generated.

Tag 8 -**Select a date to publish:** Type in any date from the date created to 5 years from that date. (The publish date defaults to 6 months from the date the collection was established.)

After changing the date(s) click on the update button  to register all the date changes in MorphBank.


Tag 9 -**Add a new annotation:** Clicking on **Add** will take the user to the **Add Annotation**




The screenshot shows a web form titled "Image Record [70756-Cynips douglasi]". It features a "Type of Annotation" dropdown menu set to "Determination". Below this is a "Related Annotations" table with columns for "Taxonomic Name", "Taxon Author", "Prefix", "Suffix", "A", "D", and "S". The table contains one entry for "Cynips douglasi" with a "Temporary TSN Name" and "none" for both prefix and suffix. The "A", "D", and "S" columns contain icons: a checkmark, a circle with a dot, and a circle with a dot respectively. Below the table is a "Determination Annotation" section with a "Determination Action" dropdown set to "New Taxon" and a "Determination Data Fields" section with a "Prefix" and "Suffix" dropdown menu, both set to "None".

screen (Figure 4) where the user can add an additional annotation to the selected object. Directions for this process are located later in this chapter.

Tag 10 - **Delete an annotation:** The last column in the annotation

manager is the delete column. To delete an annotation, click on the delete  icon. A confirmation message will appear prior to completing the delete. (This option is available only if the annotation is not yet published.)

Tag 11 -**Update button:** All alterations on the annotation manager page (date to publish changes) must be registered to become permanent. To register changes, click on the update  button.

Types of Annotations

- **Determination:** This is the most complex of the annotation types and is designed to offer biologist the ability to remotely collaborate on the determination (assignment of a taxonomic name); and to offer the ability to supply additional details concerning the taxonomic name associated with a specimen. When **Determination** is selected as the annotation type, additional field options will be available:
 1. Determination annotation will give users the ability to view and respond to a list of determination annotations that are related to the current object.
 2. Users can choose to comment on the previous determinations, select a new taxonomic name from the ITIS database, or add a new taxon.
 3. Users are required to provide MorphBank with the source of the identification (defaults to the name of the logged in user) and resources used in making this determination annotation.

An annotation title, comments and date to publish are the remaining required fields in this option. (Details for this annotation type are located in the [Add Annotations](#) documentation below.)

Note: Even though the image was selected for annotation, it is really the associated specimen that is linked to the determination annotation. For example, if two users create a determination annotation using two different images from the same specimen, when the determination annotations are viewed for that specimen, both will be seen as related annotations. If a determination annotation is written for a collection of images there will be an identical determination annotation record written for each specimen in the collection.

- **General:** This annotation type is used to add general comments about an image or collection of images. The required fields in this option include an annotation title, general comments and date to publish (The publish date defaults to 6 months from the date the collection was established.) (Details for this annotation type are located in the [Add Annotations](#) documentation below.)
- **Legacy:** General and legacy annotations differ only in the source of the annotation. Data in a legacy annotation was previously generated and stored elsewhere prior to the inclusion in MorphBank. As in a general annotation, a legacy annotation is used to add general comments about an image or collection of images. The required fields in this option include an annotation title, general comments and date to publish (The publish date defaults to 6 months from the date the collection was established.) (Details for this annotation type are located in the [Add Annotations](#) documentation below.)


- **XML:** This option allows the user to upload an XML document into the MorphBank database and use it as a general annotation. All other fields match the general and legacy annotations. The required fields include an annotation title, general comments and date to publish (The publish date defaults to 6 months from the date the collection was established.) The XML document is limited in size to 64K. (Details for this annotation type are located in the **Add Annotations** documentation below.)

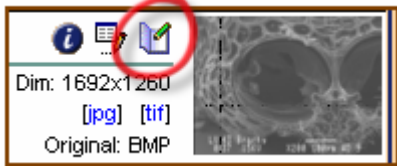
Add Annotations

New single (one at a time) or mass (multiple) annotations are added through the **Add Annotation** or **Mass Annotation** screens.

Adding new single annotations:

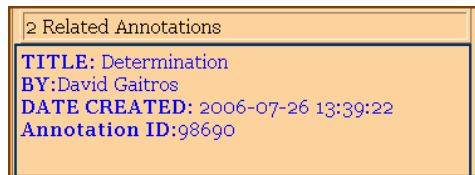
Single annotations are created through **Browse-Images**, through the results of a **Search**, through an existing annotation (i.e. annotation manager, annotation-show, related annotations etc.) or through a **Collection** (i.e. browse-collection, collection, collection-show etc.)

To reach the add annotation screen, logged in users can select the annotation icon  located beside the thumbnail image of the record to be annotated as seen in **Browse-Images** or through the results of a search.




other annotation screens by clicking is a selectable **Related Annotations** blue highlighted type). The related screen contains a dropdown menu annotations.

The **Add Annotation** screen can also be accessed in **Annotation Manager** by selecting the **Add** column, or in



anywhere there area (noted by annotation for add

Users can also access the single annotation process from **My Collection**, by checking one image in a collection (check the box in the lower left side of the image). Then click on **Annotate Checked Objects** or select the annotation icon  located beside the thumbnail image of the record to be annotated



Add Annotation
 User: Emerald Systems (Logout)
 Group: PBEP/Canada Academic/Canada (Coordinator)

Image Record [70758-Cynips douglasi]

Type of Annotation * Determination

Related Annotations

	Taxonomic Name	Taxon Author	Prefix	Suffix	A	D	S
<input type="radio"/>	Cynips douglasi	Temporary TSN Name	aff	none	1	0	1
<input type="radio"/>	Cynips douglasi	Temporary TSN Name	none	none	1	0	1

Determination Annotation

Determination Data Fields

Determination Action * Give different name - choose name below

New Taxon

Prefix None

Suffix None

Materials used in Id Image

Source of Identification *

Resources used in Identification *

Common Annotation Fields

Title * Determination

Comments *

Image Label:

X-Coord

Y-Coord

Date To Publish (YYYY/MM/DD) * 2007/02/01

Submit Return

*-Required

Figure 4 Add Annotation

This figure contains test data

All required fields are followed by an *.

- Type of annotation: (Required field) The default selection for this field is **Determination**. The other options of **General**, **Legacy**, and **XML** are selected from the drop-down list.
- Related annotations: (available only with the annotation type of **Determination** selected.) The user can select from a list of previously submitted, related determination annotations for that image (or related images) To select the related annotation, click on the radio button to the left of the taxonomic name. This field also contains a history of the previous annotations (author, prefix/suffix, A (agree with taxon name), D (disagree with taxon name), S (number of specimen(s) associated with this determination and collection of images).

Attributes of **Related Annotations** in the list for a single determination annotation:

1. All annotations in the list have the same specimen (specimen id)
2. All annotations in the list must be determination annotations

3. Included in the related annotations list is the initial determination placed in the specimen record.




This means that all of the images associated with a single specimen will have the same related annotations visible in a determination annotation.

- Determination action: (**Required field** that is available only with the annotation type of **Determination** selected.) and choose to agree, disagree, or agree with qualification (to agree with the taxon but not with a listed prefix or suffix.)

Agree: The user must choose a previous determination using the radio buttons to the left of the related annotation. An annotation record will be added that agrees with that taxonomic name, prefix and suffix.

Disagree: The user must choose a previous determination using the radio buttons to the left of the related annotation. An annotation record will be added that disagrees with that taxonomic name, prefix and suffix.

Qualify lowest rank: The user must choose a previous determination using the radio buttons to the left of the related annotation. Additionally, the user will have the ability to qualify the taxon with a prefix and/or suffix. (These appear only after the qualify option is selected) The combination of taxonomic name/prefix/suffix must be unique (if there is a duplicate, an **Agree** annotation will be added).

- New taxon: (available only with the annotation type of **Determination** selected.) If no related annotation was chosen from the list, the user has the option of selecting a new Taxon name from a list. To insure accuracy, taxonomic names need to be selected  from the **Taxonomic Selection Screen**. Traverse through the levels  until the appropriate scientific name is found. Then click the select icon , it will automatically direct the user back to the add annotation screen and the appropriate name will be filled in.

If a new taxon name needs to be added select the **Add new Taxon** button that is visible from the family level. The **Add TSN** screen will popup. (This option is only available for authorized users.) For complete instructions on this process see the [ITIS, Add New Taxon](#) section of this manual.

Note: Great care must be taken when adding new taxon names to the local copy of the database. New names must be accurate and accepted in the biological community. Adding a new taxon name commits the user to the responsibility of submitting a change to the Department of Agriculture <http://www.itis.usda.gov/>.







- Prefix/suffix: (available only with the annotation type of **Determination** selected; and only available if user chose to agree with qualification or chose a new taxon name.) Users can choose a prefix or suffix from the appropriate drop-down list to qualify their determination action.


Prefix options include:

<ul style="list-style-type: none"> • <i>None</i> 	<ul style="list-style-type: none"> • <i>Forsan (perhaps)</i>
<ul style="list-style-type: none"> • <i>Not</i> 	<ul style="list-style-type: none"> • <i>Near (close to)</i>
<ul style="list-style-type: none"> • <i>Aft (akin to)</i> 	<ul style="list-style-type: none"> • <i>Of lowest rank</i>
<ul style="list-style-type: none"> • <i>Cf (compare with)</i> 	<ul style="list-style-type: none"> • <i>? (questionable)</i>

Suffix options include:

<ul style="list-style-type: none">• <i>None</i>	<ul style="list-style-type: none">• <i>Senso Stricto (in the narrow sense)</i>
<ul style="list-style-type: none">• <i>Senso latu (in the broad sense)</i>	<ul style="list-style-type: none">• <i>Of lowest rank</i>

-  Materials used in id: (available only with the annotation type of **Determination** selected.) Indicate the materials examined to formulate this determination annotation by selecting an option from the drop-down list.
-  Source of identification: (**Required field** that is available only with the annotation type of **Determination** selected.) Enter the name of the person who made the determination. The default for this option is the logged in user. The name can be changed if the annotation is being made on behalf of someone else.
-  Resources used in identification: (**Required field** that is available only with annotation type of **Determination** selected.) Indicate the resources used to support the determination annotation. This is a free text entry for information such as citations of literature or expert opinion.
-  Title: (**Required field**) Click on this field to change or enter a title for the annotation. The default title is **Determination** for a determination annotation. For other types of annotations enter an appropriate descriptive title.
-  Comments: (**Required field**) Enter comments to support the annotation or comments that might aid other users to understand the particulars of this annotation, or add any other information that might be useful to keep with the annotation. Examples: explain why the specimen was identified with the particular taxon, comment on an image marker placement etc.
-  Image label: When annotating a single image, the user has the option of identifying a location on the image to associate a pointer and label (If annotating a group of images this option will not be available).



To add a marker to the image, select the  beside the **Image Label** field. The current image will display. Click on the screen (do not drag) where the point of the marker is to be located. To reposition the marker, click on the screen in the new location. The old marker will be replaced by a new marker.

The marker color can be selected. Click the radio button next to the desired color (choices are red [default], blue, yellow and green).

To add a label to the marker, type the label in the **Annotation Label** field provided on the screen.



When the image has been marked and labeled, select submit. The screen returns to the add annotation screen. If a marker label was added, it will show up in the **Image Label** box. As long as the annotation is not yet published, a submitted marker can be changed through edit annotation.

-  X/Y coordinates: This field will display automatically after a marker has been placed on the image. It is not suggested that the coordinates be manually changed by the user. The location of the marker on the image is represented as a percentage (%) of pixels from the left of the image (x) and from the top of the image(y).
-  Date to publish: (**Required field**) Type in any date from the date created to 5 years from that date. (The publish date defaults to 6 months from the date the collection was established.)

- Submit/Return: Select **Submit** to upload the annotation data to MorphBank and go back to the place where the annotation was initiated or select **Return** to go back to that screen without submitting any data.

Adding new mass annotations:

A user with a login account can annotate a group of images called a “mass annotation”. Mass annotations can be made through any area in MorphBank that accesses collections i.e. **Browse-Collection, Collection, Collection-Show**, etc. By selecting all or any subset of a group of images, a user can request to annotate that collection by calling the add annotation screen and entering the data. This will cause an annotation record to be added for each individual image selected. Additionally, if the annotation type was a determination, then a **Determination Annotation** record will also be added or created through the **Annotation Show** function (fig).

To access annotations through **Browse-Collection**, locate the collection to annotate. Click on **Edit**  then proceed as directed below for **My Collection**.



Users can access the mass annotation process from **My Collection**, by checking images in a collection (check the box in the lower left side of the image). Then click on **Annotate Checked Objects**. If only one image is selected to be annotated, the user will be directed make a single annotation.



Mass Annotation
 User: [username] (logout)
 Group: [group]

About Browse Search Tools Help

1 4 Images of 26 in Collection 98765 [Cynips douglasi Coll.]

2 Image Record: [70779] Image Record: [70776] Image Record: [70765] Image Record: [70756]

Type of Annotation * Determination

Related Annotations

Taxonomic Name	Taxon Author	Prefix	Suffix	A	D	S
<input type="radio"/> Cynips	Temporary TSN Name	aff	sensu stricto	1	0	1
<input type="radio"/> Cynips	Temporary TSN Name	aff	none	1	0	1
<input type="radio"/> Cynips douglasi	Temporary TSN Name	none	none	4	0	4

3 Determination Annotation

Determination Data Fields

Determination Action * Give different name - choose name below

New Taxon

Prefix None

Suffix None

Materials used in Id Image

Source of Identification * David Gaitros

Resources used in Identification *

Common Annotation Fields

Title * Determination

Comments * Annotation related to the following images: 70779, 70776, 70765, 70756 of Collection id [98765]

Date To Publish (YYYY/MM/DD) * 2007/02/07

Submit Return

*-Required

Figure 5 Mass Annotation
 This figure contains test data

All required fields are followed by an *.

Tag 1– Mass annotation heading: This displays the collection id and name that the mass annotation was initiated from as well as the number of images that were selected to annotate from the collection.

Tag 2– Image thumbnails: This list of thumbnails represents the images that are included in this mass annotation. The list will scroll as needed to display all included images.

Tag 3- Related annotations: (available only with the annotation type of **Determination** selected.) This list will contain all specimens associated with the images contained in tag 2 above.

Taxonomic name - represents the lowest level taxonomic name of the specimen.

Taxon Author - Author of the taxonomic name from the ITIS database.

History – This contains the historic data relating to prefix(s)/suffix(s) and totals regarding previous annotations associated with this determination. A (agree with taxon name), D (disagree with taxon name), S (number of specimen(s) with that taxonomic name and collection of images).

The instructions for the remaining fields contained on the mass annotation page can be found in the [Add Single Annotation](#) section of this manual.

Note: The reference to image markers and labels on the add annotations page are not available for mass annotations.

Edit: Annotation

Edit Annotations contains the previously entered annotation data that can be edited by the owner (only available if the annotation is not yet published.) Make note that the type of annotation can not be altered.

Edit Annotation is accessed through the **Annotation Manager** by selecting **Tools/ Annotation Manager** or **Tools/Edit Annotation**.



The **Edit Annotation** screen will come up when the user clicks on the title of the annotation that is to be edited.

Annotation Manager

Clicking on the annotation title will bring up the Edit Annotation screen.

Id*	Title	Type Annotation	Object Id	Object	Date Created	Publication Date	Add	Delete
100020	Test Determination	Determination	70756	Image	2006-07-28	2006-09-01	add	×
104133	Test of Annotation Location	General	99697	Image	2006-08-15	2006-08-14	add	×
104134	Test 2 of Annotation Tool Arrow Location	General	99697	Image	2006-08-15	2007-02-14	add	×
104135	Arrow Location, Lower Left	General	99697	Image	2006-08-15	2007-02-14	add	×
104136	Test of Arrow, Upper Left	General	99697	Image	2006-08-15	2007-02-14	add	×
104137	Determination	Determination	99697	Image	2006-08-16	2007-02-15	add	×
104138	Determination	Determination	70756	Image	2006-08-17	2007-02-16	add	×
104139	Determination	Determination	99697	Image	2006-08-18	2007-02-17	add	×
104140	Determination	Determination	99697	Image	2006-08-18	2007-02-17	add	×
104141	Annotation without label on Arrow	General	99697	Image	2006-08-23	2007-02-22	add	×

Update

Figure 6 Edit Annotation

This Figure Contains Test Data

Annotation Record of Image [70756-Cynips douglasi]

Type of Annotation Determination

Related Annotations

Taxonomic Name	Taxon Author	Prefix	Suffix	A	D	S
<input checked="" type="radio"/> Cynips douglasi	Temporary TSN Name	none	none	2	1	1

Determination Annotation

Determination Data Fields

Determination Action * Agree - choose name above

New Taxon Cynips douglasi ✓

Prefix None

Suffix None

Materials used in Id Image

Source of Identification * Visual Exam

Resources used in Identification * Expert opinion of Dr. Philip S. Aspinwall

Common Annotation Fields

Title * Test Determination

Comments * This is an example of a single determination from the collection screen. No previous determinations were made of this specimen and as such the Specimen record shows up as the initial determination. Update of Comments

Image Label: Throat ✓

X-Coord 51

Y-Coord 28

Date To Publish (YYYY/MM/DD) * 2006/09/01

Update Return

*-Required

Figure 7 Edit Annotation

This Figure Contains Test Data

The information included on the **Edit Annotation** screen reflects all the previous data that was included on the original annotation. To edit the information on this page, click on the appropriate area to highlight the data and type in or select the corrected information. Make note that the type of annotation can not be changed; however, if the annotation has not been published, it can be deleted entirely and reentered under the proper type. Help in filling out the data fields on this page can be obtained in

[Add Annotations](#) located in this manual.

Annotations Record Show:

This is an example of an annotation record page displayed from the **Annotation Manager / Id** (first column). MorphBank **Single Show** is an efficient way to display large amounts of information. For complete documentation on single show refer to [MorphBank Show](#) in the **Information Linking** section of this manual.

Annotation Manager					
Id*	Title	Type Annotation	Object Id	Object	D
100020	Test Determination	Determination	70756	Image	26

Annotation Record: [100020] Title = Test Determination

Contributed By: David Gaitros

Date Contributed: 07-28-2006

Last Modified: 07-28-2006

Publish Date: 09-01-2006

Specimen Id: [68530]

Sex: Female

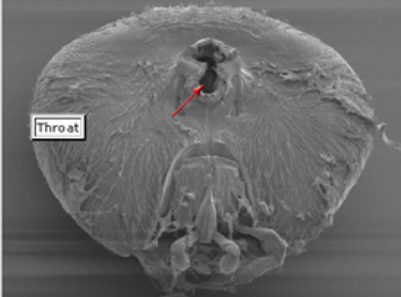
Collector: Johan Liljeblad & Fredrik Ronquist

Species Name: Cynips douglasi

Object Id: [70756]

Object Type: Image

Type of Annotation: Determination



Comments

This is an example of a single determination from the collection screen. No previous determinations were made of this specimen and as such the Specimen record shows up as the initial determination. Update of Comments

Related Annotations to this image

TITLE: Test Determination
TYPE ANNOTATION: Determination
BY: David Gaitros
DATE CREATED: 2006-07-28 13:44:01
RELATED ANNOTATIONS OF ID: [100020]
SINGLE SHOW OF ANNOTATION ID: [100020]

TITLE: General Annotation
TYPE ANNOTATION: General
BY: David Gaitros
DATE CREATED: 2006-08-21 14:36:14
RELATED ANNOTATIONS OF ID: [104905]
SINGLE SHOW OF ANNOTATION ID: [104905]

Related Annotations1

Determination Data

Specimen Id: [68530]

Taxonomic Serial Number: [999000435] Cynips douglasi

Taxonomic Name: []

Prefix: [none]

Suffix: [none]

Type Determination: [agree]

Source of Id: [David Gaitros]

Resources used in Id: [Expert opinion of Dr. Fredrik Ronquist]

Materials used in Id: [Image]

Taxonomic Name	Taxon Author	Prefix	Suffix	A	D	S
Cynips	Temporary TSN Name	none	none	2	0	1

Select to access Related Annotations page

Figure 8 Single Show-Annotation

This Example Contains Test Data

In **Annotations Record Show**, the user is presented with a list of all annotations associated with the object to include all annotations related to the image and specimen.

Clicking on this section will bring up the **Related Annotations** page which contains all the tools needed for a user to research annotations associated with the current annotation record.

Use the Related Annotation page to:

- View a scrollable list of all images related to the annotation of the same specimen, images with the same taxonomic name, images with the same view, and images in collections to which the current image belongs.
- Email an annotation to another party for viewing.
- View related image, specimen or view data. This option utilizes the MorphBank Show option to display a full set of information on the image data, the specimen data, or data about the view associated with the image.
- Add a new annotation to the current image by calling the single add annotation screen. Or sort the current list of related annotations

Related Annotations:

Related Annotations contains all the tools needed for a user to research annotations associated with the current annotation record.

The **Related Annotation** page is designed to display all of the information associated with a particular annotation and display links to detailed data on the specimen, image, view, locality, and determination. Additionally, **Related Annotations** permits the user to view the image in more detail using a commercial image viewer product called the **FSI Viewer from Neptune Labs**.

Since any single object within MorphBank may have several related objects, this screen displays some of those relationships. The user can, by selecting the related image drop-down menu, display other images related to the current image, specimen of the image, species, images with the same view, or all of the images in collections where the current image is also contained.

The screenshot displays the MorphBank interface for 'Related Annotations'. At the top, a navigation bar includes 'About', 'Browse', 'Search', 'Tools', and 'Help'. Below this, a secondary navigation bar contains 'Related Images', 'Mail', 'View', 'Image', and 'Annotation'. The main content area features a large SEM image of a specimen's head, with a red arrow pointing to a 'Throat' label. A callout '1' points to the top left navigation area. Callout '2' points to the image title 'Image Record [Cynips douglasi]'. Callout '3' points to the image viewer controls. Callout '4' points to the top right navigation area. Callout '5' points to the image viewer scroll bar. Callout '6' points to the annotation details section, which includes fields for 'Contributed by: David Galloway', 'Date Contributed: 07-28-2006', 'Publish Date: 09-01-2006', 'Title: Test Determination', 'Species Name: Cynips douglasi', 'Related Specimen: [68530]', 'Type Annotation: Determination', and 'Determination Record: [100020]'. Below the metadata is a 'Return to Admin' button, with callout '7' pointing to it. At the bottom, there are navigation links for 'Contact Us', 'Copyright', and 'About MorphBank'. Callout '8' points to the image viewer zoom controls.

Figure 9 Related Annotations

This Figure Contains Test Data



Tag 1: Clicking on **Related Images** will display a drop-down list for the user to select which category of **Related Annotations** to display. The related images will display in the list at the bottom of the page.

with a MorphBank URL that will allow the recipient to view the using the **MorphBank Show** feature.

Tag 2: The **Mail** option allows users to the annotation URL to any valid email for viewing with an accompanying user message. A sent email contains the text



email address supplied along image



Tag 3: The View drop-down box displays a selection choice of record data types that can be displayed on the screen.

Tag 4: This option brings up the commercial image viewer product called the **FSI Viewer from Neptune Labs.** gives the user many more viewing

Complete instructions for this viewer can be found in the FSI manual located in this manual,



This viewer options. Viewer



Tag 5: Use this option to add an annotation or sort the onscreen list of related annotations by title, author, or date. The previous order of related annotations and collection images are maintained

Tag 6: List of related annotations. This list is also a hot-link that allows the user to display that annotation data on the current web page. Select to reveal that annotation

Tag 7: List of related images. This is where the images from tag 1 above are deposited. Additionally, clicking on the thumbnail images of the related images will display related annotations associated with that image.

Tag 8: . The image is shown in a larger format than normally seen in the rest of MorphBank. The image displays any designated annotation marker and label, (overlaid arrow and label). Clicking on the image brings up the image viewer which allows the image to be viewed in more detail using a commercial image viewer product called the **FSI Viewer** from Neptune Labs.

APPENDIX D

EXAMPLE: MORPHBANK XML IMAGE ANNOTATION SCHEMA

```
<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">
    <sequence>
      <element name="point" type="point-type" minOccurs="2"
        maxOccurs="2" />
    </sequence>
  </complexType>
</schema>
```

```

</complexType>
<complexType name="loc-type">
  <choice>
    <element name="rectangle" type="rec-type"/>
  </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>
<element name="annotation" type="annotation-type">
</schema>

```

BIBLIOGRAPHY

1. [ASK04] Akraiva, G. Stamou, G., and Kollias, S. "Semantic Association of Multimedia Document Descriptions through Fuzzy Relational Algebra and Fuzzy Reasoning", IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, volume 34, number 2, March 2004
2. [AlRuAn03] Alexander, L. Runyan, A., and Anderson, V., "Taxonomic Data Working Group, Darwin Core 2", <http://darwincore.calcademy.org>, 2003
3. [ACLE03] Aloisio, G., Cafaro, M., Lezzi, D., van Engelen, R. "Secure Web Services with Globus GSI and gSoap", Proceedings of the EUOPAR, 2003
4. [ArLi03] Arenas, M. and Libkin L., "An Information-Theoretic Approach to Normal Forms for Relational and XML Data", Proceedings of the Twenty-Second ACM SIGMOD-SIGIACT-SIIGART Symposium on Principles of Database Systems, 2003, pages 15-26
5. [Cant04] Canton, C., "An Experience in Building an Ontology-Driven Image Database for Biologists", Standards and Ontologies for Functional Genomics Conference, October 2004, University of Pennsylvania School of Medicine
6. [ChJoBe99] Chandrasekaran, B., Josephson, J., and Benjamins, V., "What are Ontologies and Why Do We Need Them?", IEEE Intelligent Systems, Jan/Feb 1999, pages 20-26
7. [CFKST01] Chervenak, A., Foster, I., Kesselmen, C., Salisbury, C., and Tuecke, S., "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets", Journal of Network and Computer Applications, volume 23, 2001, pages 187-200
8. [Cdpw02] Ciravengna, F., Dingli, A, Petrelli, D, and Wilks, Y., "User-System Cooperation in Document Annotation Based on Information Extraction", In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02, Springer Verlag, 2002
9. [Dcw03] Dingli, a., Ciravegna, F., and Wilks, Y., "Automatic Semantic Annotation using Unsupervised Information Extraction and Integration", Workshop on Knowledge Markup and Semantic Annotation, KCAP03, 2003
10. [ElNa04] Elmasri, R. and Navathe, S., "Fundamentals of Database Systems, Fourth Edition", Addison-Wesley, Reading, Massachusetts, 2004
11. [FoKe98] Foster, I., and Kesselmann, C., "The Grid: Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishing, San Fansisco, CA, 1998
12. [Foster02] Foster, I., Vöckler, J., Wilder, M., and Zhao, Y., "Chimera, A Virtual Data System for Representing, Querying, and Automating Data Derivation", 14th International Conference on Scientific and Statistical Database Management, Edinburgh Scotland, July 2002

13. [Gait04] Gaitros, D. “*Common Errors in Large Software Development Programs*”, Crosstalk, Journal of Defense Software Engineering, March 2004, pages 21-25
14. [GRRJ05] Gaitros, D., Riccardi, G., Ronquist, F., Jammigumpula, N., and Blanco, W. “*MorphBank, The Development of a General Purpose Bioinformatics Database*”, International Conference on Internet Computing (ICOMP’05), June 2005, pages 31-37
15. [Gait06] Gaitros, D., Ronquist, F., Dean, Riccardi, G., Mast, A., Jorgensen, P. “*Morphbank, The Requirements and Implementation of a Digital Image Phylogenetic Database*”, Taxonomic Databases Working Group, St. Petersburg, Russia, October 2006,
16. [GZMRR06] Gaitros, D., Zhang, W., Mast, A., Riccardi, G., Ronquist, F. “*A Biodiversity Semantic Associative Annotation Tool*”, *The 2006 International Conference on Internet Computer and Conference on Computer Games Development*”, CSREA Press, Las Vegas, Nv June 2006, Pages 29-35
17. [Gruber93] Gruber, T. R., “*Towards principles for the design of ontologies used for knowledge sharing*”, Presented at the Padua workshop on Formal Ontology, March 1993
18. [GuWiGu03] Guthrie, D., Wilks, Y., Guthrie, L.,”*Using Semantic Annotations to Cluster Lexical Relationships*”, Johns Hopkins University Summer Workshop 2003, 2003, http://www.clsp.jhu.edu/ws2003/documents/postworkshops/guthrie_final_rpt.pdf
19. [Hass96] Haas, L., Kossmann, D., Wimmers, E., Yang, J., ”*An Optimizer for Heterogeneous Systems with Non-Standard Data Search Capabilities*”, In Special Issue on Processing for Non-Standard Data, IEEE Data Engineering Bulletin 19(4), Dec 1996, pages 37-43
20. [Hala01] Halascheck-Weiner, C., Hunter, J., Simou, N., Smith, J., and Tzouvaras, V., ”*Image Annotation on the Semantic Web*”, http://www.w3.org/2001/sw/BestPractices/MM/image_annotation.html, Jan 2001
21. [HaSt03] Handschuh, H., Staab, S., “*Annotation of the Shallow and the Deep Web*”, Annotation of the Semantic Web, Volume 96 Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, 2003, pages 24-45
22. [Hsm01] Handschuh, S., Staab, S., and Maedche, A., “*CREAM, Creating Relational Metadata with a Component-base Ontology Drive Framework*”, In Proceedings of K-CAP 2001, Victoria, BC, Canada, October 2001
23. [Holli04] Hollink, L., Schreiber, G., Wielemaker, J., Wielinga, B. ,”*Semantic Annotation of Image Collections*”, Workshop on Knowledge Markup and Semantic Annotation, KCAP03, 2004, <http://www.cs.vu.nl/guss/papers/hollink03.pdf>
24. [Jörg96] Jörgenson, C.,”*Indexing Images: Testing an Image Description Template*”, ASUS 1996 Annual Conference Proceedings, 1996
25. [Kim02] Kim, H., “*Predicting how Ontologies for the Semantic Web Will Evolve*”, Communications of the ACM, volume 45, February 2002, pages 48-54

26. [KoMaSc05] Korica, P., Maurer, H., Scerbakov, N. "Extending Annotations to Make them Truly Valuable", World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (ELEAN) 2005, 2005
27. [KoSw03] Kouvunen, M. and Swick R., "Collaboration through Annotation on the Semantic Web", Annotation for the Semantic Web, Volume 96 Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam, 2003, pages 46-60
28. [Kreg01] Kreger, H, IBM Support Group, "Web Services Conceptual Architecture", <http://us.ibm.com>, 2001
29. [LeGl01] Leonard, T., Glaser, H., "Large Scale Acquisition and Maintenance from the Web Without Source Access", In Siegfried Handschuh, Rose Dieng-Kuntz, and Steffen Staab, editor, Proceedings Workshop 4, Knowledge Markup and Semantic Annotation, K-CAP 2002, 2001
30. [Lewi02] Lewis, S., Searle, S., Harris, H., Gibson, S., Iyer, V., Richter, J., Wiel, C., Bayraktaroglu, L., Birney, E., Crosby, M., Karminker, J., Matthews B., Prochnik, S., Smith, C., Tupy, J., Rubin, G., Misra, S., Mungall, C., Clamp, M., "Apollo: A Sequence Annotation editor", Genome Biology 2002, 3(12):research0082, 2002
31. [LiRo98] Liljeblad, J. and Ronquist, R., "A Phylogenetic Analysis of Higher-Level Gall Wasp Relationships(Hymenoptera: Cynipidae)", Systematic Entomology, volume 23, 1998, pages 229-252
32. [Lije06] Liljeblad, J., Ronquist, F., Nieves-Aldrey, J. L., Fontal-Cazalla, F. M., Ros-Farré, P., Gaitros, D. & Pujade-Villar, J. (in manuscript)., "Morphological Phylogenetics in the Information Age: Relationships among Oak Gall Wasps and Their Closest Relatives (Hymenoptera: Cynipidae).", Zootaxa, 2006
33. [Liu04] Liu, C., Nguyen, T., Zhang, R., Yao, F., Zhu, Z., Gershon, E., "SNP Information Mining Pipeline (SIMP) for Complex Disease Studies", 53rd Annual Meeting of the American Society of Human Genetics, Los Angeles, CA, 4-8 November, 2004, AM J Human Genet 73(5):421. abstract #1470
34. [Liu03] Liu, C., Bonner, T., Nguyen, T., Lyons, J., Christian, S., Gershon, E., "DNannotator: Annotation Software Tool Kit for Regional Genome Sequences.", Nucleic Acids Res. 2003. volume 31(13), pages 3729-3735
35. [Marsh97] Marshall, P., "Annotations: From Paper Books to Digital Library", In Proceedings of the ACM Digital Libraries 97 Conference, Philadelphia, PA, July 1997
36. [MaSmSz05] Martin, S., Smith D., and Szekely, B. "LSID (Life Science Identifier) Project", 2005, <http://lsid.sourceforge.net>
37. [RoGaRi06] Mast, A., Ronquist, F. Gaitros, D., Riccardi, G., "MorphBank, An Open Web Repository for Biological Images", Botany 2006 Workshop, California State University, Chico , Ca, August 2006.

38. [Meng04] Meng, C., “*Biological Information Standards*”, Bulletin of the American Society for Information Science and Technology”, 2004
39. [MiKn05] Michleson, M., and Knoblock, C., “Semantic Annotation of Unstructured and Ungrammatical Text”, Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh Scotland, July 2005
40. [Morel96] Morell, V., “*TreeBASE: the roots of phylogeny*”, Science, 273:569, 1996
41. [Moun00] Mount, S., “Genomic Sequence, Splicing, and Gene Annotation”, American Journal of Human Genetics, October 2000, Volume 67(4), pages 788-792
42. [Myers04] Myers, J., “*Scientific Annotation Middleware, SAM04*”, <http://collaboratory.emsl.phl.gov>, 2004
43. [MCGS03] Myers, J., Chappell, A., elder, M., Geist, A., and Schwidder, J., “*Re-Integrating the Research Record*”, IEEE Computing and Science Engineering, May, 2003
44. [MyGe03] Myers, J., Geist, A., “*Scientific Annotation Middleware, SAM*”, Office of Science, United States Department of Energy, <http://www.csm.ornl.gov/DOE/mics2003/>, 2003
45. [Nakh03] Nakheh, L., Miranker, D., Barbancon, F., Piel, W., and Donoghue, M , “*Requirements of Phylogenetic Databases*”, 3rd IEEE Symposium on Bioinformatics and Bioengineering”, 2003
46. [NG00] Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD, Lasky SR, Baliga NS, Thorsson V, Sbrogna J, Swartzell S, Weir D, Hall J, Dahl TA, Welti R, Goo YA, Leithauser B, Keller K, Cruz R, Danson MJ, Hough DW, Maddocks DG, Jablonski PE, Krebs MP, Angevine CM, Dale H, Isenbarger TA, Peck RF, Pohlschroder M, Spudich JL, Jung KW, Alam M, Freitas T, Hou S, Daniels CJ, Dennis PP, Omer AD, Ehardt H, Lowe TM, Liang P, Riley M, Hood L, DasSarma S. Genome sequence of a genetically tractable and extremely halophilic archaeon. *PNAS* 97: 11677-12388, 2000.
47. [NGM02] Nguyen, T., Liu, C., Gershon, E., McMahon, F., “*Frequency Finder: A Multi-Source Web Application for Collection of Public Allele Frequencies of SNP Markers*”, Bioinformatics 2003
48. [Oldfi03] Oldfield, R., “*Summary of Existing Data Grids*”, <http://www.cs.dartmouth.edu/~raoldfi/gridwhite/paper.html>”, 2003
49. [Olsen04] Olsen, G., “*The Newick’s 8:45 tree format*”, http://evolution.genetics.washington.edu/phylip/newick_doc.html
50. [Onei02] O’Neill, K., “*The Web and the Grid: from e-science to e-business*”, <http://www.w3c.rl.ac.uk/EuroWeb>, 2002
51. [Parr04] Parr, C., Lee, B., Campbell, D., and Bederson, B., “*Visualization for Taxonomic and Phylogenetic Trees*”, Oxford Journal, Life Sciences, Bioinformatics, Volume 20, number 17, 2004, pages 2997-3004

52. [Peil96] Piel, W., Donoghue, M., Erickson, T., Henze, C., Rice, K., and Sanderson, M., "*TreeBase: A Relational Database of Phylogenetic Knowledge*", Society of Systematic Biologists, St. Louis, Missouri, 1996
53. [PiDoSa02] Piel, W., Donoghue, M., and Sanderson, M., "*TreeBASE: a Database of Phylogenetic Knowledge*", In J. Shimura, K. Wilkson, and D. Gordon editors, *To the interoperable "Catalog of Life" with partners – Species 2000 Asia Oceania*, pages 41-47. 2002. Research Report from the National Institute for Environmental Studies No. 171, Tsukuba, Japan
54. [Ricca01] Riccardi, G., "*Principles of Database Systems with Internet and Java Applications*", Addison-Wesley, Reading, Massachusetts, 2001.
55. [Ricca06] Riccardi, G., "*Representing and Using Phylogenetic Characters in MorphBank*", Taxonomic Data Working Group (TDWG) Annual Meeting, St. Louis, Mo, October 2006
56. [Ricca05] Riccardi, G., Gaitros, D., Jammigumpula, N., Blanco, W., van Engelen, R., Ronquist, F., "Managing Image Annotations", Data Integration Application Workshop on Data Grids, Edinburgh Scotland, August 2005
57. [RoGaMa04] Ronquist, F., Gaitros, D., Mast, A., "*MorphBank: Web Image Database Technology for Biology*", Taxonomic Database Working Group, University of Canterbury, Christchurch, New Zealand, October 2004
58. [RoSe91] Rosenthal, A., and Seigal, M., "*Flexibility and Control Requirements for Metadata Integration Systems*", WITS Workshop, 1991
59. [Sande93] Sanderson, M., Baldwin, B., Bharathan, G., Campbell, C., Ferguson, D., Porter, J., Dohlen, C., Whohciechowski, M., and Donoghue M., "*The Growth of Phylogenetic Information and the Need for a Phylogenetic Database*", *Systematic Biology*, volume 42, 1993, pages 562-568
60. [SCHEMA] The W3C RDF Schema Working Group, URL:<http://www.w3.org/TR/WE-RDF-Schema/>
61. [SDWW01] Schreiber, A., Dubbeldam, B., Weilemaker, J., Weilinga, B., "*Ontology-Based Photo Annotation*", *IEEE Intelligent Systems*, 16(3), pages 66-74, May/June 2001
62. [Shan02] Shan, H., Herbert, K., Piel, W., Shasha, D, Wang, J, "A Structured –Based Search Engine for Phylogenetic Databases", 14th International Conference on Science and Statistical Databases, Pages 7-10, 2002
63. [Shann48] Shannon, C., "*A Mathematical Theory of Communications*", *Bell Systems Technical Journal*, volume 27, 1948, pages 379-423
64. [SmMaSz05] Smith, D., Martin, S., and Szekely, B., "*LSID (life Science Identifier) Project*", 2005, <http://lsid.sourceforge.net>

65. [SpMeJa03] Spyns, P., Meersman, R, and Jarrar, M., “*Data Modeling versus Ontology Engineering*”, SIGMOD Record, December 4th, 2003, Volume 31, pages 12-17.
66. [Szala02] Szalay, A., Kunszt, T., Thakar, A., and Gray, J., ”*Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey*”, Technical Report MS-TR-((-30, Microsoft, December 2002, pages-29-35
67. [TreeBASE] “*A Database of Phylogenetic Knowledge*”, <http://www.treebase.org>
68. [VaPa99] Vasudevan, V., Palmer, M. “*On Web Annotations: Promises and Pitfalls of Current Web Infrastructure*”, 32nd Hawaii International Conference on Systems Sciences, Jan 1999
69. [Vargas02] Vargas-Vera, M., Motta, E., Dominique, J., Lanzoni, M., Stutt, A., and Ciravegna, F., “*MnM: Ontology Driven Semi-Automatic or Automatic Support for Semantic Markup*”, In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02, Springer Verlag, 2002
70. [Xie02] Xie, H., Wasserman, A., Levine, Z., Novik, A., Germiskey, V., Shosan A., and Mintz, L., ”*Large Scale Protein Annotation through Gene Ontology*”, Genome Research 12, 2002, pages785-794
71. [Zhan02] Zhang, L., “*Maximum Entropy Modeling Toolkit for Python and C++*”, <http://www.nlplab.cn/zhang/maxentoolkit.html>, 2002, pages 1-4

BIOGRAPHICAL SKETCH

The author holds a BA in Computer Science, 1977 from Southern Illinois University, and a MS in Computer Information Sciences from the Air Force Institute of Technology, 1985. The author is a retired Air Force Officer Lieutenant Colonel with over 22 years of active duty service. He is currently the Director of User Services for the Office of Technology Integration at Florida State University.