

## **ABSTRACT**

The University of Michigan Library and the University of Michigan's Museum of Natural History seek funding for a collaborative effort to enhance access to collections of the Museum of Natural History through the use of digital technology and digital library methodologies. This project has two primary goals: 1) the development of increased access to the Great Lakes region Flora and Fauna portions of the Museums collections, and, 2) the development of an extensible infrastructure for putting Natural History collections online. The project will explore, prototype and test tools for using the online resources for a variety of scholarly and educational purposes.

In this phase of the collaboration, the project will apply the resources and methods of the University of Michigan Library's well-established Digital Library Initiative to the Museum of Natural History's Fish, Fungus and Mammal collections, three collections that have already made considerable use of digital technology to organize and provide access to their content. Access will be enhanced by attention to several aspects of the collections:

1. The digitization, using best practices and existing standards, of a substantial body of material from the collections pertinent to describing and illustrating the flora and fauna of the Great Lakes region.
2. Enhancement of the existing descriptive records associated with the collection to facilitate use by a varied audience and to promote better interoperability with other related collection databases and catalogs.
3. The creation of standards and models for future data creation that are sensitive to the needs of a wide audience using online resources.
4. Integration of the collection databases to allow simultaneous cross-collection searching – so, for example, a researcher can search the databases by location and find items from all the collections associated with that location.

Also in this phase of the collaboration, the development of the infrastructure to support online search, retrieval and use of these collections will focus on enabling:

1. Cross-database searching as described above.
2. The use of data with spatial coordinates – providing access to species information through maps and location information as well as traditional descriptive records.
3. Identification and retrieval of both bibliographic and non-bibliographic types. The collections are extraordinarily heterogeneous, consisting of items that range from traditional monographs to hand-written field notes to DNA samples to skeletons. The infrastructure will be designed to accommodate this great variety of content.

Finally, this collaboration seeks to develop an online resource for multiple audiences, from primary school users to advanced scholars in the field, to the general public. The Great Lakes region focus will create a body of digital materials that capitalizes on an area of collection strength from the Museum and generates both local interest from educators and students and nationwide interest from scholars of the natural history of the Great Lakes region. The resulting online collection will feature such important data as images of type specimens from the mammals of the Great Lakes region, field notes and guides from collecting expeditions from the region, and hard-to-find monographs that contain important information for specimen identification. The interface to the collection will provide tools for searching and viewing this rich variety of museum data and will also be a centralized site for posting educational materials and for pointing to other resources related to the flora and fauna of the Great Lakes region.

## **NARRATIVE**

### **NATIONAL IMPACT**

Flora and Fauna of the Great Lakes Region : A Multimedia Digital Collection will provide an opportunity to assist Natural History museums in sharing information and to combine Library expertise in information organization and access with the rich subject knowledge and collection strengths of museums. It is indisputable that museums will struggle, perhaps for decades, to adopt and retrofit new standards to old systems for managing and storing data. Although there have been commendable efforts in various museum communities to develop standards for data description and dissemination, many problems still remain. The University of Michigan Library has developed federating methods for combining databases so as to promote integrated access and cross-database searching while at the same time preserving access to the heterogeneity and valuable subject detail of the individual databases. This federating model acknowledges and respects diversity among museums' organization and description methods, and it also facilitates conversion and migration, enhances searchability of descriptive data, provides broad public access to text and images, and drastically reduces the effort required to disseminate information. It is our expectation that the obvious benefits of these methods, as demonstrated through the resulting access system, will encourage other institutions to explore adoption and further development of the strategies used at the University of Michigan.

In addition to model methods, it is our hope that the success of this project will also suggest a model of cooperation. The project brings together disparate strengths from across the University for a common purpose. The University Library contributes established methods for organization of information and serving a varied audience. More specifically, it contributes the subject knowledge and public service expertise of the Science Library, the technological expertise of the Digital Library Initiative, including the knowledge of descriptive standards of the Library metadata specialist. The Museum of Natural History contributes highly specialized subject expertise, a sensitivity to users of all kinds (the collection curators are both scholars and teachers) and extraordinary collections of the museums.

Bio-diversity is a topic of huge interest to the K-12 community, higher education and the public. Teachers attempting to incorporate inquiry-based learning into their classroom activities often need examples of the diversity of flora and fauna to use as the raw materials for this learning. Locating these raw materials is very difficult, often requiring fairly specific knowledge to access the databases if they are accessible at all. Museum collections have been used for years in research, and more recently in education especially at the higher education levels, but use of these materials is most often accomplished when the faculty member teaching the class is a curator of collections. Opening these collections up to a wider audience is important, as these resources are exactly the kind that current teachers and students need and want to use. It is our hope that by creating an online structured search tool for working with disparate collections, this project will promote inquiry-based learning about the natural history of the Great Lakes region.

In addition, the materials in most natural history museum collections represent a very rich set of resources that augment the specimen records. Our project will include access to this rich array of materials and will describe the full range of materials available in these collections so that further digitization work can easily add to what is currently available. Creating the methods for locating materials regardless of the particular database the materials reside in means greater access to the novice museum patron, student, or teacher.

The Museum of Natural History is a collective body consisting of the University of Michigan Exhibits Museum of Natural History, the University of Michigan Herbarium, and the Museum of the Zoology. Three of the Museum collections, represented by five online databases, are targeted for this first phase of collaboration -- the Fungus, Mammal and Fish collections. The University of Michigan Fungus Collection (MICH) is one of the five largest in the Western Hemisphere, housing 280,000 specimens, 22,709 notes on specimens, 11,960 images, and 1,570 paintings and drawings. The collection is the result of nearly 100 years of research by present and former curators. The core of the collection is the thousands of collections of higher fungi that were used to describe 2,000 plus species of North American mushrooms, gasteromycetes, and false truffles in more than 34 major monographs. Most of these collections are only available from the University of

Michigan museums. The Mammal collection is the third largest university-based mammal collection in this country, and it ranks fifth or sixth in size among all US mammal collections, and includes around 150,000 specimens. This collection emphasizes not only breadth of coverage, but also the accumulation of series of specimens to document geographic and within-population variation. It contains the largest accumulation of specimens from the Great Lakes region in the world and 130 primary types. Finally, the Fish collection has recently been ranked by the American Society of Ichthyologists and Herpetologists as an International Center, second in the United States only to the Smithsonian Institution. The collection consists of over 3 million specimens from nearly 66,000 localities worldwide. Major strengths of the collection are the Great Lakes region, the American Southwest, Southeast Asia and Japan, Venezuela and Paraguay, and Mexico. Some of the collections from these regions are the most comprehensive in the world. The collection includes nearly 10,000 skeletons. Collection data include about 150,000 pages of field data, 5,000 images of specimens, and 3,000 X-ray plates. Improved public access to those collections is one of the great benefits of this project and will have great benefit public visibility and widespread appeal among students of all kinds and ages.

## **ADAPTABILITY**

This project will create, implement and document workable models that have the potential for successful, wide-scale adaptation. Moreover, the models explored and developed in this project are potentially adaptable to other kinds of scientific collection databases.

The success of the federating model has already been demonstrated in creating online access to both individual collections and to over twenty aggregated art, architecture and archaeological databases (see Appendix for Image Services URL). This project will both test and demonstrate the adaptability of that model for different kinds of data types. The proposed model combines respect for diversified management practices with federated broad access in a separate, highly reliable and efficient production system. This combination affords the museums the opportunity to begin broad dissemination without major management overhaul and still encourages their creation and adoption of standards.

The project emphasizes the use of standards such as SGML for data encoding, the World Wide Web as an Internet protocol, HTML for interface building, TIFF for storing master image files, JFIF/JPEG for image access, and ISO 9660 for archiving data to CD-ROM. None of these standards is *required* to adopt UM's approach to federated access. The use of standards and the flexibility to substitute other standards or proprietary approaches makes the model very adaptable.

In addition we encourage adaptation through sharing technical information. As an institution of higher education and as regular contributors to national digital library efforts, the University Library is vested in sharing information about system architecture wherever possible. The DTD used for encoding Image collections and the Image Services Architecture definition proposed for use in this project, among other information, has been available on the Web since inception, and will continue to be available as it evolves. Moreover, the middleware used to facilitate the online delivery of the collections is available to the educational and nonprofit entities through the University Library's Digital Library eXtension Service.

Finally, we will promote adaptation of the model through widespread dissemination of information about the project methodologies. Documentation of fundamental architecture attributes, and processes, are key to wide-spread adaptation. Processes, in particular for the conversion of materials, will be documented and appropriately summarized in publication. This may include costs per image, taking into account hardware and personnel costs, as well as information specific to the capture process for particular types of objects, for example resolution, dynamic range, and format.

## **DESIGN**

### Goals and objectives:

- 1. The digitization, using best practices and existing standards, of a substantial body of material from the collections pertinent to describing and illustrating the flora and fauna of the Great Lakes region.**

The digital conversion of the candidate materials and the creation of an access system for those materials will be undertaken by The University of Michigan's Digital Library Initiative (DLI). This Initiative was conceived in 1993 to address the issues of coherence and coordination of the University of Michigan's networked information resources in an increasingly distributed technology environment. In the seven years since, DLI has developed the necessary infrastructure and organizational support to shape this program into a nationally recognized leader. A core group with the DLI is the Digital Library Production Service – a unit providing digitization services, information retrieval system building and production level support for online scholarly collections. DLPS manages over 300 gigabytes of digital content, as well as supporting programs in content creation, conversion, and systems design.

The heterogeneous nature of the collections and of the items selected to be digitized – ranging from handwritten field notes to published guidebooks to photographs to actual specimen data – will require a number of different conversion methodologies. DLPS has well-established methods for the creation of fully processed and encoded text, bitonal page images and fully searchable OCR text, high resolution continuous tone images, flatbed scanning of originals, and flatbed scanning of film intermediaries. In addition, DLPS will use this project as an opportunity to explore appropriate digitization methods for three-dimensional objects, such as skeletons. One of the aims of this project is to determine the most appropriate methods of digital capture for the many kinds of data and materials in the museum collections. Please refer to the section on digitization methods for more extensive information on methods and standards.

## **2. Federation of the collection databases to allow simultaneous cross-collection searching**

This project seeks to provide a single unified point of access to the online collections of the Museum of Natural History, and at the same time provide an easy path to working with individual collections in their full and discipline-specific detail. The project will establish coherence among disparate collections in two ways. First, by concentrating on the flora and fauna of the Great Lakes region, the project will encourage a parallel development of a portion of several different collections; it will make all the Great Lakes region materials accessible online and will impose some common descriptive practices on those portions of the collections. As an ancillary benefit, this will provide a single location for links to other relevant collections to information about related items and events from the Exhibit Museum. Second, it will make use of the Digital Library's federating mechanisms to promote cross-collection search and retrieval of the collection databases.

Image Services, a specialized group within DLPS, has a proven track record of providing federated access to image collections which are managed in a distributed fashion; currently it provides centralized access to over twenty image collections that use a wide variety of data description schemes. Central to this model is the notion that public access must be separate from internal management of collections. Museums and visual resource collections in general, have well-established practices for collection description, but these practices do not necessarily promote broad access. Michigan acknowledges and values inherent differences among collections that have long histories of distributed management, and the federating model respects and accommodates these differences. The federating model allows museums to provide their data in any structured text format, and with any defined set of data categories, thus reducing overhead for sharing data. The model implements cross-collection access by transforming data fields into SGML elements and mapping the diverse collections to a common Document Type Definition (DTD). This transformation enhances searching of individual or multiple collections with standard tools for data description. Future improvements in cataloging practices and standard data exchange formats will only strengthen broad, generalized access to collections. The existing databases accessed through Image Services mechanisms primarily consist of fine art, architecture and archaeological materials; This project will establish a proof of concept of the federating model's ability to greatly advance the accessibility of natural history collections through Library approaches to disparate data access.

This model provides access to collection databases. Access includes search, retrieval, and display of records and images, but does not address item level record management, which includes creation, deletion, and modification. Database management is typically and more appropriately handled by separate systems under the control of the units that possess collections. For example, all three collections have powerful databases and interfaces for managing its large and varied collection. Enhanced public access to its data will be provided separately, by exporting data from the management database and importing it into Image Services access system. The advantages of this approach are:

- Management systems are not burdened by frequent public use.
- The access system is built to provide access only.
- A variety of management systems exist. Subject domain needs and budgets vary greatly. The access system accepts data from any management system.
- Standardization at the level of access provides consistent service for instructional and research needs. Production level servers, network access, security, backups, and maintenance provide high quality network delivery of images with minimal down time.

In addition to improved access through federation, all of the online materials will have available to them the full range of search capabilities offered by the DLXS XPat search engine. The XPAT engine is an SGML/XML-aware search engine that the University of Michigan has deployed with an extremely diverse set of digital library resources. XPAT is based on the search engine previously marketed by Open Text as OT5™, and sometimes referred to as “Pat” and “Pat5.0.” The University of Michigan has licensed the code in order to undertake distribution and support, as well as to add functionality. Its capabilities include simple word and phrase searching of all text, as well as a variety of advanced search possibilities, with search restrictions as made possible by the data. Each format will have available to it a variety of display and navigation tools appropriate to that format, such as panning and zooming tools for continuous tone images and multiple views. All materials will be searchable through an integrated interface that allows searching any combination of formats.

### **3. Enhancement of the existing descriptive records associated with the collection to facilitate use by a varied audience and to promote better interoperability with other related collection databases and catalogs.**

This project will enhance existing records by incorporating standard metadata forms

- Dates will be reformatted to comply with ISO 8601 format
- Common names will be added whenever they exist to formal taxonomic nomenclature to allow for search and retrieval by non-specialists.
- Where accurate information is available, latitude and longitude in tagged form will be added to location information. Where coordinates cannot be accurately derived from existing data, a lookup table will be provided that links location name(s) with the closest possible coordinate matches (see goal number 5, below).
- Best practices will be established for metadata associated with new forms of digitized objects, such as field notes and three-dimensional virtual reality reproductions.
- Selected fields from records will be mapped to the Z39.50 attribute in order to support later work in support of Z39.50 compliant access for outside users desiring to search local collections in conjunction with other Z39.50 applications.

Where existing electronic reference or authority data are available, an attempt will be made to develop tools that incorporate the above enhancements via automated processing. Where the necessary references are in print only, records may need to be hand-edited or reviewed.

We believe that this enhancement of existing metadata will facilitate use by a varied audience and promote better interoperability with other related collection databases and catalogs.

### **4. The creation of standards and models for future data creation that are sensitive to the needs of a wide audience using online resources.**

The application of the federation model to the museum databases, in combination with the enhancement of the metadata will promote varied public access to the collection materials and promulgate methods and standards which other institutions can adopt in working with their own collections.

Currently, each collection database has an organization and uses descriptive methods that are appropriate for a scholarly audience. These will remain in place and be accessible through the individual interfaces for each collection. We recognize and respect that collections generally need to preserve their individuality in terms of searching. That is, in order for a searcher to explore the specialties of a collection and its organization and documentation, it must be possible to investigate that collection alone. For example, fungi are naturally cataloged much differently than mammal skeletons or field notes from fish collecting expeditions. A searcher with a tight research focus will want to concentrate on the most appropriate collection. On the other hand, searching multiple collections of disparate content can result in new discoveries and ideas. Therefore, through a common interface, it will be possible to search across multiple image collections. While some information may be lost through the aggregation of the databases, the discovery of new kinds and combinations of pertinent material may well be an acceptable trade-off.

To promote greater and more diverse access to these collections, this project also aims to create a public/educational interface. The addition of common names and the standardization of dates and locations will provide clearer access for the general public and for students at all levels, a clarity that the educational interface can build on. This interface will build on the successes of the *Animal Diversity Web*, associated with the mammal collection and already used in biology teaching at the University of Michigan (as well as in other university, secondary and primary classes around the country) as well as serving as a natural history encyclopedia for the general public, and *Fun Facts About Fungi*, used by elementary, middle, high school, and university teachers around the globe and listed as a resource in an undergraduate general biology text. The educational interface will be shared with teachers in the Great Lakes region. It will also be permanently displayed on a kiosk in the University of Michigan Exhibits Museum of Natural History.

All of this work will help establish standards and models for future data creation. Once federated access is enabled and different audiences have access through different databases, collection managers and catalogers will have a better sense of the needs of multiple audiences and the uses those audience make of the data and can be better informed in their response to those needs and uses. Moreover, the standardization and addition of descriptive data to enhance access will create a body of material against which users needs and descriptive methods can be tested.

#### **5. The use of data with spatial coordinates – providing access to species information through maps and location information as well as traditional descriptive records.**

Geo-spatial coordinate information is invaluable to the study of species in the Great Lakes region since these data document the location where a specimen was collected or where a species lives. With location information in spatial coordinate form, scholars and learners can bring to their studies a wide variety of linked data in order to see what flora and fauna exist in particular locations, to trace the history of land use, to trace how species have moved over time, to study how climate relates to species location, or to develop environmental impact studies. Geo-spatial coordinate data also support user-friendly map-based database interfaces for locating materials of interest to K-12 and higher education students and teachers as they study “what is in their backyards”.

In order to facilitate this long-term goal of having geo-spatial coordinate information for all natural history collections, the current project will standardize location information across the databases, by assigning tagged coordinate data to identifiable locations and providing a look-up table for less accurately pinpointed locations. This standardization will employ automatic means whenever possible, but will also call for some intellectual review and correction by a metadata specialist.

In addition, we will digitize one hundred historical maps from the 1920s that summarize the distributions of mammals in the Great Lakes area as they were in the early 1800s. These maps are valuable resource in themselves. They will also serve as building block for facilitating later work with spatial data.

#### **6. Identification and retrieval of both bibliographic and non-bibliographic types.**

The Natural History collections are extraordinarily heterogeneous, consisting of items that range from traditional monographs to hand-written field notes to DNA samples to skeletons to art work. The infrastructure will be designed to accommodate this great variety of content. The system will support search and retrieval of print items that have already been cataloged, of specimen data embedded in database structured records, and of materials that have not previously been cataloged in any way, such as field notes and correspondence.

Key to this effort will be designating a data model for field notes and guides that both adequately describes these materials and provides for links to the specimen data which the guides detail. Part of the work of this project is to investigate whether any existing model, such as that developed by the Museum of Vertebrate Zoology at UC Berkeley, is sufficient or whether a new model needs to be developed.

When working with the physical collections or the existing databases that describe them, it is difficult to easily access information about many types of material at one time and certainly difficult to look at collections in combination. In the online environment that allows searching many types of material, it will be possible, for example to simultaneously look at a historical map detailing the fauna of the Michigan area, read field notes on the behavior of some of those fauna, and to look at images of skeletons. By providing access to this rich variety within and across collections, this collaboration seeks to engage users in new patterns of inquiry.

### Work Done to Date

#### *Drawing on Work at the University of Michigan*

The work done for this project draws substantively on work done at the University of Michigan partner organizations, and also will build upon the foundation of work undertaken with Natural History data in other initiatives. The University Library brings to the project several years of experience in improving and centralizing access to a number of different types of data systems as well as kinds of data. It also has done extensive work in developing an infrastructure to support online delivery of scholarly data and has in place robust authentication and authorization mechanisms, production level support of all hardware and a staff with a wide range of technical expertise. The three candidate collections for this project have all benefited from substantial work done to enhance access to their data using digital technology. All three collections have electronic catalogs and the curators have provided online access to information about the collections. The Fungus specimen database is online as are a number of supporting materials about the collections. A companion website, "Fun Facts about Fungi" is an educational site for the general public that offers inquiry-driven, active use of the Fungus Collection online database, a geographic information system, images of fungi, keys for identification, and species descriptions. The curator of the Mammal collection has worked extensively with materials from that collection to create the extensive Animal Diversity Web. The Fish collection catalog is searchable, in combination with the catalogs of three other large fish collections, through the Biodiversity and Biological Collections Web Server. This project will draw upon the partners' strong foundation of technical knowledge and engagement with technological developments in information delivery.

The project will also build upon digitization efforts that are already underway. The University Library has worked with the fungus collection to digitize 371 plates from fungus guides, and three hundred detailed watercolor paintings of mushrooms. As part of the work for Animal Diversity Web, thousands of images of animals in natural settings and hundred of recordings of animal sounds have been digitized and the site includes more than three hundred Quick Time Virtual Reality movies of specimen skulls.

#### *Drawing on work from outside Michigan*

The project will also capitalize on the work that other institutions have undertaken in attempting to build better data models and systems for Natural History. For example, The Museum of Vertebrate Zoology at UC Berkeley has already done substantial work on a metadata model for Natural History collections. This project will use that work to inform the metadata enhancement. The work done for this project also differs from that done by UC Berkeley in some significant

ways. While the MVZ works assumes a *new* description of the data, this project seeks to work with *existing* records and to capitalize on descriptive work that has already been done and to improve that work by applying standards such as the ISO compliant dates and standardized locality information. This project assumes that domain specific metadata standards will remain in place and will examine which standards work *across* the data sets.

Members of the Specify project at the University of Kansas have taken Berkeley's model and used it to develop software for collection management. While, as has already been mentioned, our University of Michigan project assumes independent collection management, features of the Specify model will be applied to the IMLS grant project, in particular, the use of authority files to maintain consistent normalized data. Further, we believe that the Kansas attempt to produce a model that applies to all types of collections will give us better insight into which similarities among collections offer the best possibilities for cross-collection coordination.

Considerable work had also been done to provide unified access to diverse collections through the use of the Z39.50 protocol, such as the work done to support Z39.50 access to seventeen mammal collections. While our proposed project's use of the federating model poses a different solution to unified access, work can also be done to make use of Z39.50 in relation to other institutions' projects. We believe that the federated model provides richer access to distributed databases than Z39.50, but we will also provide a map from our metadata fields to the Z39.50 attribute set cross-institution Z39.50 to promote and support later work on Z39.50 compliant access.

## **MANAGEMENT PLAN**

The University Library's Digital Library Initiative is a long-standing program with a strong national reputation for taking on projects, carrying them out with high technical quality, and completing the projects on time and on budget. These characteristics will be true of this project as well. DLPS will be the locus of much of the activity for this grant. Christie Stephenson, Assistance Head of DLPS, will be the project lead under the direction of John Price-Wilkin, head of DLPS. Ms. Stephenson has been involved in several large and notable digital library and museums projects. She established the Digital Image Center at the University of Virginia in 1993 and served as Project Director for the Museum Educational Site Licensing Project.

The University of Michigan Natural History Museums will be the partner on this grant. The 5 museums that make up the Natural History Museums, 3 of which will participate in this project, have a long history of excellence in research, database management and curatorship, as well as a strong exhibit program through the Exhibits Museum. The directors of the University Library and the Natural History Museums are fully supportive of this proposal.

The University's Division of Research Development and Administration will work with the project team to ensure that financial management of the grant is done appropriately.

## **BUDGET**

### Direct Costs

#### *Salary and Wages (Permanent Staff)*

Funding for the following work is being requested from IMLS: Metadata Specialist, Programmer, and some of the digitization. The permanent staff members listed in the budget are working on other projects. If this grant is awarded, the projects will either be done by other staff or delayed.

#### *Salary and Wages (Temporary Staff Hired for Project)*

Graduate students will be used for field note reviewing and annotating for taxonomic name and location at a cost of \$15 per hour. Students for photography work will cost approximately \$12/hour. The Collection Implementer will be hired for the duration of the project.

### *Fringe Benefits*

Benefits are calculated at 28% for permanent and temporary positions that are salaried positions and 8% for hourly positions. The University of Michigan covers fringe benefits for UM salary costs. Fringe benefits for IMLS salary costs are requested from IMLS.

### *Materials, Supplies, and Equipment*

Materials listed in this category are supplies for digitizing: CD-ROMs and film. In addition, because of the number of items being handled, a separate CD-burning workstation will be required to complete the items within the grant period. A Macintosh G4 is requested for this use. A digital camera is requested to add to the ones already in use at DLPS due to the large amount of content to be photographed in the two years of the project. Two workstations are requested for the Exhibit Museum kiosk.

### *Services*

The DLPS OCR service components have been quantified and the rates approved by the University of Michigan Cost Reimbursement Office. The cost for OCR is \$0.10 per page.

### Indirect Costs

The University of Michigan considers this project a non-research project and therefore, the indirect costs are calculated at 30% of direct costs. Indirect costs are requested from IMLS for IMLS direct costs only.

### No Costs Incurred

No costs are being incurred for this project for servers, disk storage, software or network access. These items are being provided by DLPS. Additionally, no cost are being incurred for up to 5% of the curators' time spent selecting materials, advising on image quality, and providing requirements for the interface.

## **PERSONNEL**

Christie Stephenson, Assistant Head of DLPS, will provide overall project management and coordination for the project. The programmer for this project will be John Weise. John has been responsible for creating the federated model digital images collection at the University of Michigan. Digitization will be done by Sally Bjork, Rashmi Nikore, and Bill Hall from DLPS. Data loading and CD burning will be done by Nita Patterson. The Metadata Specialist will be Judy Ahronheim, who has a wide range of experience in cataloging and the use of metadata for digital collections.

The person hired as the Collection Implementer for the duration of the project, will design the interface, evaluate the end product as described in the Evaluation section, and will do whatever is necessary to bring the collection online and make it usable for students and teachers.

Graduate Students with subject expertise related to the collections that can assist in identification and description of materials.

Curators will be selecting materials for digitization, advising on accuracy of capture and quality of images, and overseeing graduate students as they review field notes for taxonomic names and geographical locations. They will also provide requirements for the educational interface to the collections. They are: Dr. Robert D. Fogel (Curator, Herbarium (Fungi) and Professor, Biology Department); Dr. Philip A. Myers (Associate Curator, Museum of Zoology (Mammals) and Associate Professor, Biology Department); and Dr. William L. Fink (Curator, Museum of Zoology (Fish) and Professor, Biology Department). Ms. Amy Harris (Development Officer, Exhibit Museum) will be working with the project team to develop the Exhibit Museums kiosk.

## **EVALUATION**

The project's success will to some extent be measured by the degree of interest generated from various communities and the volume of use from those communities. We will be working to promote use of the system in local classrooms, to engage metadata specialists in discussion of our metadata standards, and to attract use by research scholars. Several

measures will be taken to test whether the project improves access to the collections through a viable infrastructure. Improvement in access will be measured by the increased volume of data made available online, by the volume of user interaction with the interface, and by qualitative measures. Details are listed below.

The success of the digitization portion of this proposal can be measured in a number of quantitative ways:

- Number of objects digitized.
- Number of types of objects made accessible.
- Number of images made accessible.
- Number of image pixels made accessible.
- Number of bytes delivered.
- Servers are always accessible

In addition, the digitization of these items will make possible user studies that can assess the effectiveness of the digital objects as surrogates for the originals.

The success of federation of databases and of data enhancement will be determined by more qualitative methods such as feedback from end-users and from participating museum directors and curators. This feedback will be collected through one-on-one interviews and focus groups, as well as observation in participating classrooms and online and paper surveys. These methods will also help determine whether the user interface meets the requirements of collection managers and end users.

It is important to note that one of the primary activities of the collection implementer appointed to this project will be evaluation and assessment. This will include collecting ongoing informal feedback as the system is developed and the collections go online and more formal assessment in the second year. The more formal assessment will use a variety of tools and resources such as log analysis, surveys, focus groups, and groups of students from courses in the School of Information. The feedback generated from this assessment will inform the next rounds of development of this project.

## **DISSEMINATION**

Dissemination of the project's methods and results will proceed on a number of fronts. As the project is underway, we will establish a project web site on which is posted the original project proposal, the plan of work, periodic progress reports and links to the results of our efforts, as well as to the web sites of related projects.

The greatest means of dissemination will be the public availability and improved intellectual access to the data. The system that results from this project will ensure widespread, continuous access to the materials, with a high level of free, unrestricted access to the search and display of the materials.

Information about the bibliographic data types will also be available in other venues. While most of the contents of project collections are not bibliographic in nature, for those that are, records will be contributed, either directly into OCLC or via OCLC's CORC project. Records will also be created and shared for web interfaces created by the project.

In addition, the project will be discussed through papers and publications in both library and natural science venues. For libraries, results can be discussed and demonstrated at library conferences, shared through Digital Library Federation forums and published in journals such as DLib. Reports will also be submitted to society newsletters such as the Mycological Society of America's Inoculum and education journals such as The Science Teacher, The American Biology Teacher, and Education Leadership.

## **CONTRIBUTIONS**

Cost-sharing for this project comes from both the Digital Library Initiative and from the Museums. The Natural History Museums have spent considerable time and funds in cataloging specimen data, in beginning to capture images digitally, and in experimenting with public access interfaces to the data in their collections. The faculty on this project, who are

curators at the Museums in addition to their teaching in the Biology Department, also have long experience in using materials from these collections for education and research and are contributing 5% of their time to this project.

## **INFORMATION ACCESS**

– see Sustainability and Dissemination sections

## **DIGITIZATION DETAILS**

### Material to be digitized (Details in Appendix A)

Each of the collection curators selected a range of materials to be digitized to amplify the specimen records in their respective databases. They include a broad range of data types including original watercolors, field notes, and major monographs documenting various species. Efforts have been made to select material providing additional documentation where the Michigan collection holds the holotype and to fully document species representing the Great Lakes region.

### Uses of/Demand for material in digital form

Please see National Impact section for discussion of demand for materials in digital form.

### Copyright issues and other potential restrictions

Most materials selected for digitization are free of copyright restriction. The Curator of the Fungus Collection has applied to the copyright holder for the 11 monographs for permission to make those texts available on the web. Permission has been granted for 11 of the monographs. The Fungus Collection also holds the rights to the watercolors, slides, and 9,000 black-and-white images.

### Technical standards

Wherever they exist, established standards for digitization, description and the creation of retrieval systems will be used. Where no standards exist, the project will adopt the established best practices of the Library and Museum communities. In cases where there is no consensus on best practice we will provide careful documentation of our methods and rationale for choosing those methods. Please see Appendix B for full details of technical standards

## **SUSTAINABILITY**

The Natural History Museums have a long history of stewardship. Maintenance of research collections, including ancillary materials like images, notes, field books, etc., is a normal curatorial function. Digital information is also an important part of cataloging and tracking collections. These activities will continue as part of the usual course of business. DLPS will continue to digitize collections from the Museums as funding and/or capacity are available. The interface to the collections will be maintained by DLPS. Additionally, the servers, software, hardware, and network connectivity will continue to be provided by DLPS for access to these collections.

**Budget Notes**

Permanent Staff and Temporary Salaried Staff: All staff salaries are on the budget spreadsheet. Year 2 uses a 4.5% increase in salary for regular (non-hourly) staff. The percent effort for digitization is based on the time it takes to digitize a particular kind of resource and the number of those resources. The number of items being digitized is included in Appendix D.

For digitization, the following was used as the basis for how much could be done in an hour, and therefore, based on the how many of a particular type we are digitizing, what percent of an FTE is required.

| Type of Material                | Number per Hour |
|---------------------------------|-----------------|
| 4 x 5 photography plus scanning | 5               |
| 4 x 5 scanning                  | 8               |
| Bitonal scanning                | 150             |
| Flatbed scanning                | 6               |
| Grayscale scanning              | 50              |
| Kontron                         | 8               |
| Preparation and quality control | 100             |
| Slides                          | 12              |
| Data loading (per CD)           | 4               |
| CD Burning (per CD)             | 4               |

Our plan is that 75% of the digitizing will be done in Year 1, 25% in Year 2.

Graduate students are to be paid \$15/hour. We plan to hire a number of students in the natural sciences to do this work. No one student will be working enough hours to require paying tuition. Other students for QuickTime VR work will be paid at \$12/hour. These are prevailing rates for the quality of work we need.

Fringe Benefits: Fringe benefits of 28% for permanent or temporary professional staff and 8% for hourly staff have been applied to all salary figures.

Travel: Travel of \$4,000 per year has been added to the budget as required in the IMLS solicitation document.

Materials, Supplies, and Equipment: CDs are priced at \$1.50 each. Film at \$4.50 each. Our plan is that 75% of the digitizing and therefore, 75% of the materials and supplies budget will be used in Year 1, 25% in Year 2. Because of the large number of items to be digitized, we require an additional workstation with CD burning capabilities. This is priced as a Macintosh G4. In order to capture the skull and other images, a digital camera is requested. This item is exempt from indirect costs. In Year 2, funds for 2 workstations to serve as the Exhibit Museum kiosks for access to the databases are requested. We estimate \$2,000 each for workstations with large monitors and high resolution to most appropriately view the images.

Services: The DLPS OCR fees have been approved by the University of Michigan Cost Reimbursement Office. The resulting per page fee is \$0.10.

Indirect Costs: Indirect costs of 30% have been applied to all project costs (IMLS and UM) with the exception of the digital camera. Thirty percent (30%) is the rate that the University of Michigan has negotiated with the federal government for non-research grants.

## **Applicant Organizational Profile**

The University of Michigan libraries (the University Library and four regental libraries) is the 8th largest academic research library in North America, with over 7.1 million volumes. With 19 campus libraries and a collection of over 6 million volumes, the University Library provides services and resources to meet the needs of a diverse faculty and student body, and serves as a leader in digital library development and research. The University Library collection offers 70,000 current journals, over 30,000 online images, subscriptions to hundreds of newspapers, electronic access to nearly 50 bibliographic resources, 24-hour access library facilities, and over 16 million pages of locally-hosted and developed online resources. Noteworthy collections within the University Library include: the Labadie Collection of social protest literature; the American Music Collection; first editions of Darwin, Newton, Galileo, and others; the largest collection of papyri in the western hemisphere; and extensive collections of the works of Milton, Dryden, Yeats, Carlyle, Dickens, Faulkner, and others.

In addition to a rich collection, the University Library is a campus leader in information resource development, creation, and training. Reference staff answer over 300,000 questions each year, and more than 61,000 interlibrary loan requests are filled annually. Through its Faculty Exploratory and Knowledge Navigation Center, the University Library offers training for faculty in incorporating technology into classroom and research activities, and offers one-on-one assistance in using the latest information technologies.

Within the Library, The Digital Library Initiative is involved in all aspects of designing, producing and maintaining digital library services and content. In 1996, the Digital Library Production Service was formed to provide ongoing development and support of digital library content and to provide a clearly articulated framework for production support and future project activity. In 1998, Digital Library Program Development was created in order to support the developmental stages of digital projects and services. Together these two organizations, working with shared tools and resources, are responsible for the development, operation and maintenance of digital collections, including SGML text collections, journal images, museum images, and numeric/spatial data collections. The Digital Library continues to evolve, so that in addition to making systems to *access* the content of the digital library, it has the capabilities to *create* things that go into digital libraries in most of those areas. It also does considerable work behind the scenes on the things that makes it possible to do either or both of those things, for example, designing authentication and authorization mechanisms that are consistent with offering up digital content to licensed audiences; managing servers that store nearly a terabyte of material in a way that is consistent with building/serving digital libraries; and applying programming techniques and mechanisms to creation activities to facilitate putting digital objects online or storing them effectively.

## APPENDIX A – Items to be digitized

| <b>REQUESTED</b>                | <b>description</b>               | <b>method</b>                              | <b>format</b>                   | <b>number of pages/ images</b> |
|---------------------------------|----------------------------------|--|---------------------------------|--------------------------------|
| Mammals                         | Maps                             | Photograph -4 x 5 negative, Imacon scanner | 4 x 5 photography plus scanning | 200                            |
| <b>total 4x5 photo and scan</b> |                                  |  |                                 | <b>200</b>                     |
| Fungus (to be requested)        | Plates from Hygrophorus          | Scan negatives - 4 x 5 film scanner        | 4 x 5 scanning                  | 126                            |
| Fungus (to be requested)        | Plates from 4 titles             | Scan negatives - 4 x 5 film scanner        | 4 x 5 scanning                  | 293                            |
| Mammals                         | Images - black & white negatives | Scan negatives - 4 x 5 film scanner        | 4 x 5 scanning                  | 400                            |
| <b>total 4x5 scanning</b>       |                                  |  |                                 | <b>819</b>                     |
| Fungus (to be requested)        | Hygrophorus monograph            | Bitonal Scan                               | bitonal                         | 416                            |
| Fungus (to be requested)        | Galerina                         | Bitonal Scan                               | bitonal                         | 384                            |
| Fungus (to be requested)        | Crepidotus                       | Bitonal Scan                               | bitonal                         | 168                            |
| Fungus (to be requested)        | Pholiota                         | Bitonal Scan                               | bitonal                         | 402                            |
| Fungus (to be requested)        | puffballs of Michigan            | Bitonal Scan                               | bitonal                         | 130                            |
| <b>total bitonal</b>            |                                  |  |                                 | <b>1,500</b>                   |
| Fish                            | Prints                           | Flatbed scan                               | flatbed                         | 150                            |
| Mammals                         | Black & white prints             | Flatbed scan                               | flatbed                         | 400                            |
| <b>total flatbed</b>            |                                  |  |                                 | <b>550</b>                     |
| Fish                            | Field books                      | Grayscale flatbed scanning                 | grayscale                       | 4,800                          |
| Mammals                         | Field notes                      | Grayscale flatbed scan                     | grayscale                       | 4,000                          |
| <b>total grayscale</b>          |                                  |  |                                 | <b>8,800</b>                   |
| Fish                            | Specimens                        | Direct digital capture                     | kontron                         | 35                             |
| Mammals                         | Specimens (2,200)                | Direct digital capture                     | kontron                         | 8,800                          |
| <b>total kontron</b>            |                                  |  |                                 | <b>8,835</b>                   |
| Fungus (to be requested)        | Hygrophorus monograph            | OCR  | ocr                             | 416                            |
| Fungus (to be requested)        | Galerina,                        | OCR  | ocr                             | 384                            |
| Fungus (to be requested)        | Crepidotus,                      | OCR  | ocr                             | 168                            |
| Fungus (to be requested)        | Pholiota,                        | OCR  | ocr                             | 402                            |
| Fungus (to be requested)        | puffballs of Michigan            | OCR  | ocr                             | 130                            |
| <b>total ocr</b>                |                                  |  |                                 | <b>1500</b>                    |
| Fish                            | Slides (fish and their habitats) | Slide scanner                              | slides                          | 700                            |
| Mammals                         | Slides                           | Slide scanner                              | slides                          | 200                            |
| <b>total slides</b>             |                                  |  |                                 | <b>900</b>                     |
| <b>TOTAL REQUESTED (ITEMS)</b>  |                                  |  |                                 | <b>23,104</b>                  |

| <b>COST SHARE</b>               | <b>description</b>         | <b>method</b>                       | <b>format</b>  | <b>number of pages/ images</b> |
|---------------------------------|----------------------------|-------------------------------------|----------------|--------------------------------|
| Fungus                          | Lacterius monograph plates | Scan negatives - 4 x 5 film scanner | 4 x 5 scanning | 154                            |
| Fungus                          | Plates from 5 titles above | Scan negatives - 4 x 5 film scanner | 4 x 5 scanning | 525                            |
| <b>total 4x5 scanning</b>       |                            |                                     |                | <b>679</b>                     |
| Fungus                          | Lacterius monograph        | Bitonal Scan                        | bitonal        | 850                            |
| Fungus                          | Rhizopogon monograph       | Bitonal Scan                        | bitonal        | 175                            |
| Fungus                          | Hebeloma monograph         | Bitonal Scan                        | bitonal        | 209                            |
| Fungus                          | Mycena monograph           | Bitonal Scan                        | bitonal        | 550                            |
| Fungus                          | Psathyrella monograph      | Bitonal Scan                        | bitonal        | 438                            |
| Fungus                          | Boletes monograph          | Bitonal Scan                        | bitonal        | 450                            |
| <b>total bitonal</b>            |                            |                                     |                | <b>2,672</b>                   |
| Fungus                          | Lacterius monograph        | OCR                                 | ocr            | 850                            |
| Fungus                          | Rhizopogon monograph       | OCR                                 | ocr            | 175                            |
| Fungus                          | Hebeloma monograph         | OCR                                 | ocr            | 209                            |
| Fungus                          | Mycena monograph           | OCR                                 | ocr            | 550                            |
| Fungus                          | Psathyrella monograph      | OCR                                 | ocr            | 438                            |
| Fungus                          | Boletes monograph          | OCR                                 | ocr            | 450                            |
| <b>total ocr</b>                |                            |                                     |                | <b>2,672</b>                   |
| Fungus                          | 300 watercolors            | Flatbed scan                        |                | 300                            |
| <b>total flatbed</b>            |                            |                                     |                | <b>300</b>                     |
| Fungus                          | 35mm slides                | scan                                |                | 1,700                          |
| <b>total scan</b>               |                            |                                     |                | <b>1,700</b>                   |
| <b>TOTAL COST SHARE (ITEMS)</b> |                            |                                     |                | <b>8,023</b>                   |

## **APPENDIX B – Technical standards**

### **Image Capture: Methods, formats, image quality and quality control**

DLPS's Image Services will undertake the digitization of the various photographic formats and manuscript pages of field notes selected for inclusion in the project. They will employ appropriate conversion methods from among the following options. Current hardware is noted, though all DLPS hardware is on a three-year replacement schedule and may be upgraded during the term of the grant.

- grayscale imaging using a flatbed scanner: suitable for newspaper clippings and manuscript pages that are not visually distinctive enough to merit color scanning. These items will be done with a high-speed grayscale scanner with limited tonal range (Fujitsu M3096GX);
- continuous tone imaging using a flatbed scanner: suitable for flat materials such as photographs or manuscripts that are distinctive in part for their visual attributes, such as color, and which require high quality conversion methods (Agfa DuoScan);
- continuous tone imaging using a digital camera: suitable for materials that can not be captured on a flatbed scanner, and that require high quality color conversion methods (Kontron ProgRes 3012);
- continuous tone color imaging using 4x5 film as an intermediate form for scanning with a transparency scanner: suitable for exceptionally small or exceptionally large objects, manuscripts, and photographs that are otherwise out of range for the digital camera or flatbed scanner (Imacon Flextight Precision II).
- continuous tone color imaging using a 35mm slide scanner (Polaroid Sprintscan 35 Plus).

Image Services creates digital images at the highest resolution appropriate to the material and to the available conversion method. Conversion methods include digital photography (Kontron), flatbed scanning of originals, and flatbed scanning of film intermediaries. Master images are stored as TIFF files and written to gold CD-ROMs using ISO 9660 specifications. Kodak color and grayscale targets are included in scans of original materials. Scanner settings are stored on CD-ROM with the images. Image Services controls and monitors quality with a variety of methods including hardware color calibration, color imaging targets, image file format integrity, and visual inspection. Methods and the extent to which they are applied vary by project based on appropriateness.

Manuscript and photographic images will be stored on and served from a Sun Ultra 450 with dual 300 MHz processors, using the Solaris 7 operating system. DLPS Image Services uses MrSID from LizardTech (wavelet compression) for server storage and JPEG/JFIF format and compression for delivery to the user. Image processing is done with Image Magick. The digital images will be available in a range of resolutions. Typical viewing sizes include thumbnail (184 x 113 pixels), small (368 x 226 pixels), medium (737 x 452) pixels, and large (1475 x 905) pixels and include panning and zooming tools that allow a high level of investigation and study. In addition, they will have available to them other existing Image Services functionalities, such as comparative search and display. Images will be integrated into the unified project Web presence, where they can be searched and displayed in combination with textual materials.

### **Text Scanning and OCR**

The monographs selected for digitization in the proposed project will undergo a simple and automatic conversion process, following procedures laid down for the University Library's Making of America4 project. Project staff in DLPS will prepare materials, making every reasonable effort to ensure each volume is complete before conversion. Project staff will note pagination structure, and special features, such as title pages, tables of contents, and indexes will be highlighted. Staff will enter this basic structural metadata into a database, a copy of which will be sent to the scanner operator with the volumes to be converted. Project staff will use a flatbed raster scanner to scan all pages as bitonal (one bit) images (one page per image file) at a resolution of 600 dpi. In some cases, an automatic document handler can be used for the page imaging. Scanning specifications developed by DLPS and the Library's Preservation Division for the Making of America4 and other large scale digital reformatting projects will be applied to this project as well.

Quality Control. Project staff will inspect 5% of the converted images to ensure completeness, legibility, and placement of the images. This level of random sampling is already used in the University Library conversion efforts, and is a

statistically valid method that ensures a high level of compliance to our standards. When all images inspected on the random sampling basis pass quality control requirements, the entire batch of files is accepted. Should the sampling technique indicate even one unacceptable image, 100% of the images in that batch will be reviewed and sub-standard images replaced. A second sampling and quality control inspection will follow; if all images in the sample inspected meet quality control guidelines, the batch will be accepted. If not, further re-scan and re-inspection will be done until the staff are satisfied by the quality of the images captured. Staff will then write the images to CD-ROM and pass them on for post-image capture processing. Experience has shown that the single largest quality issue is skew of the digital image.

Post-capture processing. DLPS staff will process the page image files to generate OCR and simple SGML to enable search and navigation. SGML headers for the monographs are created from the MARC records for those volumes. This method is used in University Library text conversions and has proven reliable and cost-effective.

Text delivery system TIFF page images and the simple SGML files will be put online in the DLPS production environment. In order to deliver the TIFF page images to the end user, they are transformed on-the-fly to GIF images, using Tif2Gif software.

### **Descriptive Metadata**

Descriptive metadata for published and manuscript-like materials being digitized will include TEI headers. All metadata imported into the access system will be coded in SGML with a project-specific DTD. Where standard vocabularies are used, they will be tagged with identifiers. All dates will be converted to ISO standard form, and location coordinates will be tagged in a fashion that permits export into standard GIS utilities. Selected metadata from records will be mapped to Z39.50 attributes to permit Z39.50 compliant performance.

### **Plans for submitting collection-level descriptive records to bibliographic networks**

Most of the contents of project collections are not bibliographic in nature, for those that are, records will be contributed, either directly into OCLC or via OCLC's CORC project. Records will also be created and shared for web interfaces created by the project.

### **Plans for Preservation and Maintenance of Digital Files**

The University Library has established a range of strategies, guidelines, practices, and policies that define and support its initiatives and programs aimed at conversion to digital format. For text delivery systems, the image storage and access strategy for converted text privileges the master image by storing it online in the access system. The access system itself depends upon the presence of the master image file to deliver information to the user. The master image resides always in the most current technologies and moves forward through technologies in the University's dynamic computing environment. This avoids problems of earlier digital efforts where the critical version of the page image was resident in an off-line medium requiring continuing refreshing. The page images are stored in redundant arrays of independent disks (RAID) at level 5. The system is mirrored in a geographically separate area of campus.

In nearly all cases, DLPS maintains multiple copies of digital masters in a variety of parallel environments, a practice that we believe further contributes to the long-term viability of the digital resource. The data stored using RAID technology is written to digital linear tape (DLT) on an at least a monthly basis; in the case of SGML/XML data and bitonal image files, the data on tape are indistinguishable from the master. All data, whether SGML/XML data or images, are written to CD-ROM at the point of final quality control acceptance. SGML/XML text files and bitonal image files are stored redundantly, with the production version in a staffed, secure, climate-controlled machine room, and a secondary version on a development server in a secure, climate-controlled (but not staffed) machine room. Although we are not yet in a position to store masters of continuous tone (photographic and manuscript) image files on production or development servers, three identical copies of all derivatives from continuous image files are stored by DLPS (again, a production version, a secondary version in a second machine room, and a copy on tape). Masters of continuous tone image TIFF files are stored on gold CD-ROM in ISO 9660 format. In addition to text and image files on production servers, other files including system software are copied to digital linear tape on a frequent basis. In addition, the development version of the

system is backed up every week. In order to ensure that accidental or unauthorized changes or replacements do not occur, a high level of security and a rigorous permissions system is in place, allowing only designated staff members to alter the files. Moreover, once files are passed from production staff to technical staff, they are never passed back; in order to make further changes, the production staff must give the technical staff a new file that supersedes the previous file. Storing the data in a variety of media will help ensure future access. Also, by using industry standards for imaging and SGML/XML encoding for text, we ensure that we are able to more easily migrate data into future formats and systems.