Sample Application

2008 National Leadership Grants for Libraries

Research

University of North Texas

High-Throughput Workflow for Computer-assisted Human Parsing of Biological Specimen Label Data

Abstract

The University of North Texas's Texas Center for Digital Knowledge (TxCDK) and the Botanical Research Institute of Texas (BRIT) will conduct fundamental research with the goal of identifying how human intelligence can be combined with machine processes for effective and efficient transformation of textual museum specimen label information into high-quality machine-processible parsed data. This two-year project will advance understanding of the workflow and processes best able to increase access to and use of digitized biological collection metadata within the stakeholder communities comprised of biologists, natural history museum collections managers, biodiversity standards groups, and the library and information science community. A key challenge faced by all natural history collections is determining a transformation process that yields high-quality results in a cost- and time-efficient manner. The results of this research will yield a new workflow model for effective and efficient label data transformation, correction, and enhancement that can be replicated, adapted, and transferred to herbaria and other natural history collections.

Our study addresses this research problem: **What workflow provides for a combination of machine-assisted and human-assisted procedures to most effectively and efficiently convert textual data on specimen labels into machine-processable parsed data to ingest in a database and associate with the digitized specimen?** The goal of this project is to answer this question. The project goal will be accomplished through the following objectives:

- Identify and test machine processes for initial transformation of label data
- Identify human processes that act on the machine-transformed data to correct and enhance label data
- Develop, test, and assess user interfaces to support human processes
- Develop and test a workflow that incorporates both machine- and human-assisted procedures for effectiveness and efficiency in label data transformation and enhancement
- Assess quality of metadata resulting from machine and human processes

In addition to answering the research questions, the proposed study will produce the following deliverables:

- Tested and validated procedures and workflow for human- and machine-assisted transformation of specimen label data
- A replicable workflow model for transformation, correction, and enhancement of specimen label data
- Reports that document all results from various research activities carried out during the study
- Open source code used in the testbed (made available to community)

The results of this research will inform a new workflow model for label data processing that will have a core advantage of distributing collaboration on a large scale with tools that accelerate the ability of humans to accurately recognize and parse label data and to proof the accuracy of the work of others. Additionally, the workflow model can incorporate access to networked resources such as authority files and geo-referencing tools to enhance the use and appeal of the metadata, and thus enhance the use of digital biodiversity repositories.

Digitizing collections in a well-planned and standard way can increase use and exposure of collections to a more heterogeneous audience while simultaneously reducing physical handling and producing a permanent digital archive. We will adhere to standards for maximum interoperability of the parsed data resulting from the research, a first step in creating an environment for natural history collections to painlessly derive machine-processible, semantically-searchable metadata from their specimens and provide these to users around the world.

All project documents will be deposited on the project website and in other appropriate repositories to ensure long-term access to project deliverables. Effective outcomes for the target communities may be demonstrated by their interest in adopting code released as a project deliverable, and in adopting practices reported as most efficient.

High-Throughput Workflow for Computer-Assisted Human Parsing of Biological Specimen Label Data

Introduction

Herbaria are special natural history collections of preserved plant specimens created for scientific use. Currently there are about 3,000 herbaria in 145 countries, containing nearly 300 million specimens (Holmgren et al. 1990). Herbarium specimens are ideal natural history objects, as the plants are pressed flat and dried, and mounted on individual sheets of paper of standard size along with a label, creating a nearly two-dimensional object. They are accompanied by metadata: attached label data about the specimens themselves, including the scientific name, where they were collected and by whom and when, and who identified them. Each specimen also has other associated data, such as the name of the owning institution or collection, history of ownership, and information added during curation including geocoordinates, as well as measures of data quality (Morris 2005).

The call to create and expand digital repositories for natural history collections has been sounded for over a decade. Federal initiatives such as the National Biological Information Infrastructure (http://www.nbii.gov/portal/server.pt) along with a wide range of international projects such as the Global Biodiversity Information Facility (www.gbif.org) have increased the quantity of digitized images of a wide variety of specimens. Digitizing collections in a well-planned and standard way can increase use and exposure of collections to a more heterogeneous audience while simultaneously reducing physical handling and producing a permanent digital archive (Cohen & Rosenzweig 2006; National Science Board 2005). Digitizing the specimen is a necessary but insufficient step to provide effective access and use of the specimen. Converting the specimen metadata into machine-processible form is essential for semantic searching via search engines, distributed databases, and other data portals. A key challenge faced by all natural history collections is determining a transformation process that yields high-quality results in a cost- and time-efficient manner.

Our study's goal addresses this research problem: **What workflow provides for a combination of machine-assisted and human-assisted procedures to most effectively and efficiently convert textual data on specimen labels into machine-processible parsed data to ingest in a database and associate with the digitized specimen?** The Botanical Research Institute of Texas (BRIT) and the Texas Center for Digital Knowledge (TxCDK) propose to study how machines and humans can assist each other to yield high-quality and efficiently transformed specimen label data; a resulting workflow model will be evaluated in a testbed implementation. The central focus of the proposed research, however, is the workflow processes for the transformation of the label data. The testbed will use existing digitized specimens and associated metadata.

Research questions to be addressed are:

- RQ1: To what extent can machine-processes accurately transform label data from a test set of specimen labels that represents variation in label types, quality, and other characteristics (e.g., handwritten versus typescript)?
- RQ2: Which human processes can be incorporated into a robust workflow to further transform, correct, and enhance label data?
- RQ3: What user interfaces are most effective and suitable to the tasks and users in supporting human processes in the workflow?

The results of this research will inform a new workflow model for label data processing that will have a core advantage of distributing collaboration on a large scale with tools that accelerate the ability of humans to accurately recognize and parse label data and to proof the accuracy of the work of others. Additionally, the workflow model can incorporate access to networked resources such as authority files and geo-referencing tools to enhance the use and appeal of the metadata, and thus enhance the use of biodiversity repositories. In addition to answering the research questions, the proposed study will produce the following deliverables:

- Tested and validated procedures and workflow for human- and machine-assisted transformation of specimen label data
- A replicable workflow model for transformation, correction, and enhancement of specimen label data

- Reports that document all results from various research activities carried out during the study
- Open source code used in the testbed (made available to community)

The results of this research will yield a new workflow model for effective and efficient label data transformation, correction, and enhancement that can be replicated, adapted, and transferred to herbaria and other natural history collections. We believe the benefits to the broader natural history community will be transformative.

Assessment of Need

Herbarium specimens are vouchers—verified material that proves a plant species existed when and where it was collected—and are used as comparative material to identify and classify plants. Since the world's herbaria hold a record of plants spanning over 250 years, these collections are a priceless resource (Funk & Morin 2000; Funk 2002). However, most herbaria are physically accessed most often by professional and student botanists, slightly less so by those in other fields of biology and ecology, and infrequently by natural resource managers, governmental agencies, and the general public. There are estimated to be 95 million herbarium specimens in approximately 620 herbaria in the U.S. (Rabeler & Macklin 2006; Holmgren et al. 1990). About five percent of these have been databased (Rabeler & Macklin 2006; R. Beaman pers. comm.), and fewer have been digitally imaged. Many of these "databases" are not online, or are not searchable through federated portals such as the Global Biodiversity Information Facility, and therefore are invisible to the average information seeker using familiar search engines.

Although the standard format of a herbarium specimen lends itself well to digital imaging, digitizing the specimen is a necessary but insufficient step to provide effective access and use of the specimen data. As "virtual herbaria" (i.e. online herbarium databases) and the software used to create them proliferate, no adequate process exists to efficiently enable all those institutions without digitized legacy data to join the virtual herbarium community and provide their data to the world. The laborious process of manual keystroking required to parse the correct parts of the label data into appropriate fields without error has received little attention; such manual processing is the most costly step in terms of staff time and expense (for training, actual data entry, and checking/cleaning the data).

At one time there were high hopes for the capabilities of optical character recognition (OCR) software to conduct this work without human intervention and create machine-processable data from digital images. However, while the kinds of data included on and associated with a herbarium specimen are fairly standard, the labels themselves are products of individual plant collectors spanning 250 years. The placement of data fields and the explicitness of data provided vary widely, creating great difficulties for attempts at automatic parsing. The most significant issue, however, is that the majority of labels were not produced in a format that is easily machine-readable. This issue is compounded since specimens with non-OCRable, handwritten labels are often the most valuable; these older specimens can tell us the most about human effects on the Earth's vegetation over the last 250 years, including the movement of invasive species, and the loss of endangered species over time. The older specimens may be the most valuable in studies of global climate change, since flowering, fruiting, and leafing-related events are all recorded on these dated vouchers. Older specimens may also represent the final record of existence for rare species collected in habitats no longer intact. Finally, the specimens themselves are fragile and endangered by frequent handling and light exposure, and the need for these would be substantially reduced by digitization, thus prolonging their lifespan.

The BRIT Herbarium holds over one million plant specimens from around the globe. A survey was made of the complete holdings of one genus, *Artemisia* (sagebrushes and wormwoods), in the Asteraceae (Sunflower plant family). *Artemisia* represented an average holding for BRIT in terms of size (1179 specimens, or slightly over one cabinet-full), range of localities (worldwide but mostly North America and Europe) and ages of specimens (1805-2007). Only 41% of the *Artemisia* specimen labels were found to be easily machine-readable with off-the-shelf OCR software. These specimens were generally North American in origin and collected after 1950. The remaining 59% of specimen labels when processed through OCR resulted in text containing numerous errors (34%) or were handwritten and impossible to digitize without human processing (25%). Figure 1

presents a sample of the variation in the specimen labels and indicates the challenges to machine-only processes for transformation.

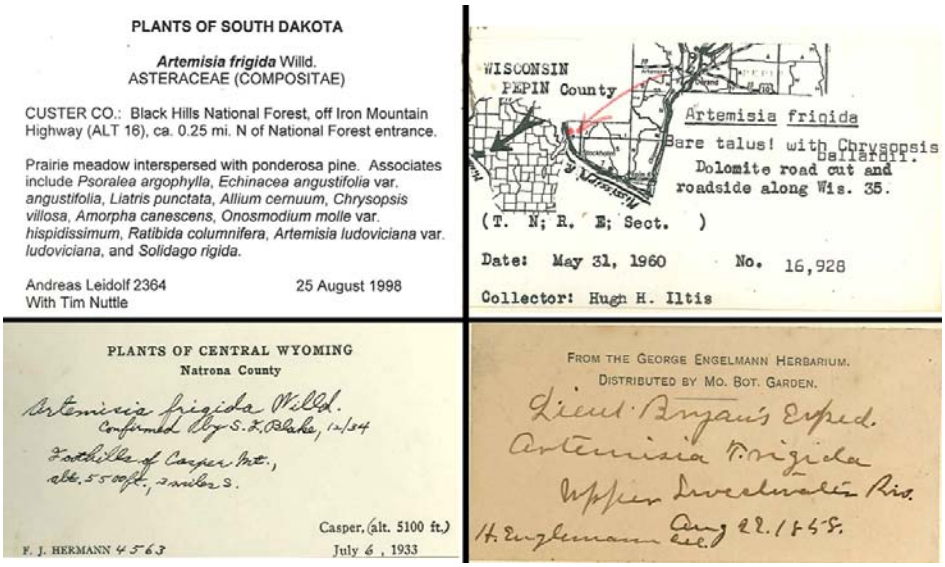

Figure 1. Typical herbarium specimen labels for *Artemisia frigida* from 1998, 1960, 1933, and 1858.

Converting the specimen label data into machine-processable form is an essential step in sharing biodiversity data with the wider world of users. A key challenge is determining a transformation process that yields high-quality results in a cost- and time-efficient manner.

National Impact and Intended Results

The need to digitize natural history collections so the data may be shared and used by a wider audience is a subject that has been at the forefront of discussion in every curatorial meeting for over a decade. Great progress has been made in the post-data entry processes that facilitate the display of data online (e.g. database software such as Specify (www.specify.org) and Brahms (http://dps.plants.ox.ac.uk/bol/home/)). However, tools to facilitate the initial digitization (especially that originating in images as the first step) have not progressed to the same extent. Nearly every institution has a backlog of non-digitized specimens, and many institutions have not yet begun to digitize their specimens. They have been left behind in the great digitization effort—especially the smaller institutions (often small university collections) with no work space for data-entry activities. A lack of familiarity with emerging data standards and a fear of producing non-standard datasets also paralyze many collections managers.

The proposed research will address these issues by providing an innovative workflow for online access in a distributed collaborative environment. Physical work space will no longer be a barrier to digitizing activities, since these activities can take place remotely and without the physical specimen present. This replicable and adaptable workflow model will incorporate human and machine processing to parse data into standard fields based on the Access to Biological Collections Data (ABCD) Schema (http://www.tdwg.org/standards/115/), an evolving comprehensive standard for the access to and exchange of data about specimens and observations, ratified by Taxonomic Data Working Group and a standard in use by providers to Global Biodiversity Information Facility (including BRIT). We will adhere to these standards for maximum interoperability of the parsed data resulting from the research, a first step in creating an environment for natural history collections to painlessly derive machine-processible, semantically-searchable metadata from their specimens and provide these to users around the world.

The most prominent effort to solve the pre-digitization issues is the ongoing development of HERBIS (“Herbis is the Erudite Recorded Botanical Information Synthesizer;” www.herbis.org), developed by collaborators at

Yale's Peabody Museum of Natural History and the University of Illinois at Urbana-Champaign Graduate School of Library and Information Science. HERBIS seeks to develop a process for creating parsed data from herbarium specimen label images using steps involving OCR, natural handwriting recognition (NHR) and natural language processing (NLP). Machine learning for maximum automation (requiring "training" of the machine learning software) is a notable focus of the HERBIS project (Beaman et al. 2006; Heidorn et al. 2007). We believe that while OCR and the machine-capable steps that follow it in HERBIS may be successful for recent specimens with well-typed and clearly organized labels, many institutions (just like BRIT) hold undigitized collections where the majority of specimens are older, with poorly-typed or handwritten labels, and these will not be efficiently transcribed or parsed through any method but pure human effort assisted by machine processes (see Supporting Document #1). The proposed research project will attempt to address the instances where OCR and NHR do not provide satisfactory results. By developing a robust workflow and testing and evaluating human interfaces in an online environment, we believe we can discover the best way to efficiently assist humans in doing their necessary part in either new input, or in correcting or proofing output from automated services, such as plain text from regular OCR applications or semantically-parsed text from HERBIS, for example.

One collaborative aspect of this project is the opportunity for three graduate students from the School of Library and Information Sciences, (UNT) of North Texas, to experience and learn more about metadata and its transformation, databases, web services, interface design and testing, and other key trends in information management. More importantly, the students will be engaged in a collaborative, interdisciplinary research project that brings them into contact with current issues in biodiversity informatics. Our findings will be of value to current and potential builders and managers of such virtual environments and help advance teaching and research in collections and metadata management, human-machine interfaces, and biodiversity informatics.

Project Design and Evaluation Plan

The proposed research project is for 24 months during which we will create a testbed to carry out a number of research activities to assess and gain an understanding of how human intelligence and machine processes can be combined in a robust workflow to transform, correct, and enhance specimen label data and metadata. The project design addresses the development of several interfaces (human and machine) to the processes and other aspects of the workflow. It also addresses the assessments of the workflow, the usability of interfaces, metrics to assess the transformation results from each workflow process, and optimal combinations of human and machine processes that yield the highest quality results with highest throughput. The evaluation plan for this research project will address assessment of the project against its goals and objectives (i.e., the success), and a preliminary indication of the impact and outcomes from the project on key stakeholders (i.e., the herbarium community).

The proposed study's goal is: **Identify how human intelligence can be combined with machine processes for effective and efficient transformation of textual museum specimen label information into high-quality machine-processible parsed data**. Answering the research questions (listed previously) and achieving the project goal will be accomplished through the following objectives:

- Identify and test machine processes for initial transformation of label data
- Identify human processes that act on the machine-transformed data to correct and enhance label data
- Develop, test, and assess user interfaces to support human processes
- Develop and test a workflow that incorporates both machine- and human-assisted procedures for effectiveness and efficiency in label data transformation and enhancement
- Assess quality of metadata resulting from machine and human processes

Three key components for conducting the research are the testbed, the workflow, and test data. The **testbed** comprises hardware, software, workflow framework, and related components. Figure 2 is a preliminary **workflow** model that indicates transformative and other processes through which the specimen images and data will move. Existing tools such as OCR engines will not be developed by the project but will be integrated into the workflow.

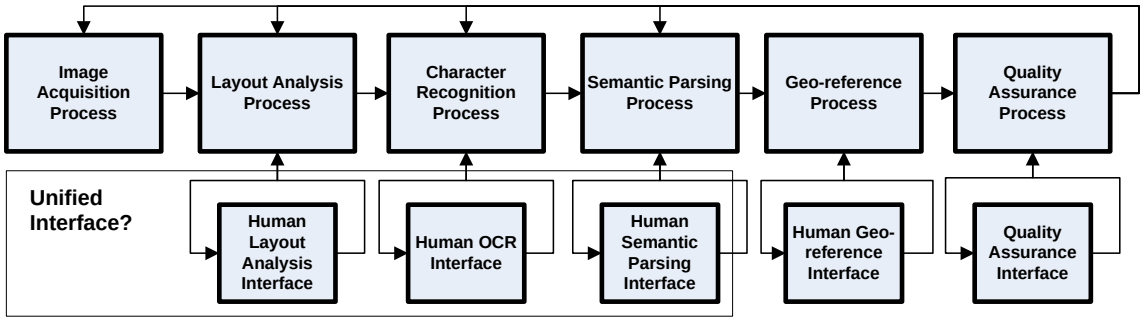

Figure 2. Transformative processes in testbed workflow

Processes in the workflow are presented in roughly in the order they may typically be executed, but the workflow can be adapted to customize the order. Research on the sequencing and interaction of the processes will be carried out.

We believe that modularity in the processes can improve the robustness and flexibility of the workflow, and we think a services-oriented architecture approach may be warranted for the testbed and workflow. Figure 3 presents several potential services that support the workflow. The project will use available services wherever possible, and only develop components of needed services to carry out the research activities. The ones indicated in bold are likely candidates for development in the project.

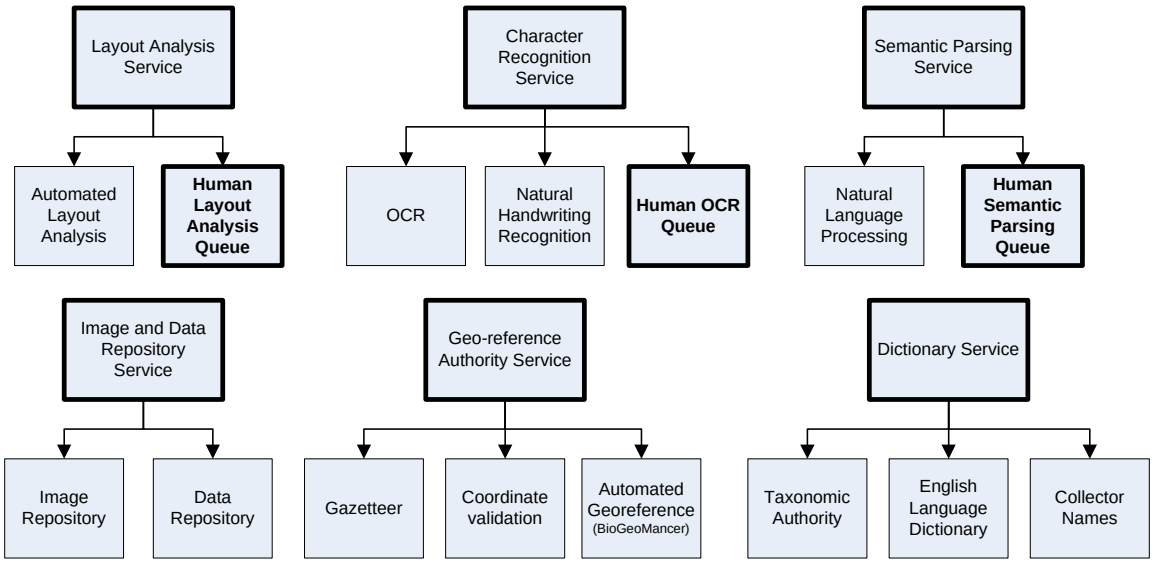

Figure 3. Services supporting the workflow

Given the proposed workflow and services, we have identified a number of potential interfaces (see Figure 2) that may be considered for supporting user interaction with label data transformation, correction, and enhancements. While these are described separately (with some repeating elements), the functions of each of the below may be provided in a unified user interface. Research and assessment will inform decisions about separate or unified interfaces.

1. Image Acquisition: Interface to bring the digitized label data into the workflow.
2. Human Layout Analysis and Image Preparation: Interface that provides tools for improving image quality (e.g., contrast correction/optimization), designating regions of interest (ROIs) and “tagging” each region as a particular type (e.g., scientific name, barcode, etc.). This will yield an improved, annotated image for all downstream workflow processes.

3. Human OCR: Interface presents the ROIs to the User along with tools for spell-checking, authority file references, etc. Allows User to enter text that they see in the ROI into standard fields. May include tools for magnifying the image, adjusting contrast for better legibility, etc.
4. Human Semantic Parsing: Interface presents the ROI and the OCRed (human or machine) text and provides tools to associate chunks of text with established, standard fields (e.g., genus, species, collector name, locality). Includes tools for interactive spellchecking, authority file reference, and possible connection to geolocating services.
5. Quality Assurance: This interface will allow the User to rate the accuracy of data, accept or reject data and queue tasks for correction. The User will have access to authority files and dictionaries as well as tools to confirm that there are no "outliers" (collection dates fall within the lifespan of collector, geo-references fall within the stated political entities, etc). The User will be able to confirm that all data comply with standards in terms of formatting and data types and will be able to check for any missing data and confirm when the full metadata record is ready for transferal into a specimen management system.

The final component of the proposed project is the **test dataset** of specimen labels to use in the research. Two primary datasets at BRIT are available for this component. The BRIT Herbarium has more than 900 type specimens that have already been digitized and databased. Type specimens are special vouchers designated as representatives of published species names. The BRIT type specimens date from 1843 to 2007 and represent a wide range and variance of label types and quality. A second dataset from BRIT's Andes to Amazon Project consists of over 1000 digitized specimens whose labels were recently database-generated. The former dataset represents a normal-case scenario of typically-encountered herbarium specimen labels; the latter represents a best-case scenario of OCRable labels with a standardized layout. The metadata extracted through the new workflows can be compared to the current metadata from these two datasets to establish the accuracy of the workflow results.

Project Phases. Additional detail on the work areas is found in the Schedule of Completion.

Startup Phase

- Work Area 0. Project Startup – Upon Notice of Award (2 Months)
  Prior to the official start date of the project (December 1, 2008), we will carry out activities to prepare for the project including: hire student research assistants, establish the project advisory board, develop a detailed project plan, establish a web presence for the project, and set up communication mechanisms for the project team and the advisory board. The detailed project plan will specify all activities, tasking, deliverables, and schedule, which will be essential for effective management of the project. The development team will begin evaluation of the various technologies that may be used in the research project, such as OCR engines, layout analysis engines, standards, and web service protocols.

Phase 1. Designing the Testbed, the Workflow, and Preparing the Test Dataset

- Work Area 1.1. The Testbed and Services (Months 1-4)
  The testbed will provide a framework in which the workflow processes will be executed and coordinates communication between the various services. The testbed will manage logging, authentication of users and services, and will present the web interfaces to users. The testbed will also provide basic interfaces to allow the researchers to monitor and modify the workflow. Various open-source workflow frameworks exist and will be assessed during the technology evaluation phase. If these do not fit our needs, a more generalized application framework will be used as the foundation for the testbed.
- Work Area 1.2. The Initial Workflow (Months 1-3)
  The first iteration of the workflow will establish the baseline for data quality and throughput by defining the optimal combination of machine processes with no human input. The participants will develop a requirements document and high level specification that will be used by developers to design the workflow framework.

- Work Area 1.3. The Test Dataset (Months 1-3)
  It is critical for the flexibility and transferability of the workflow, including human and machine processes, that the test dataset be as representative as possible of types and qualities of herbarium specimens. We will estimate the possible variation of specimen labels in the BRIT collection, assess the representation of that variation in the 900+ type specimen labels and a larger number of new database-generated labels (already in digitized form), and determine the need (if any) for additional specimens to add to the test dataset. As necessary, a small number of additional specimens and their labels may need to be digitized for the purposes of representing the widest variance in the test dataset. We anticipate the number of additional samples not to exceed 200.

Phase 2. Assessing Quality of Machine Processing of Test Dataset

- Work Area 2.1. Machine Processing of Test Dataset (Months 2-4)
  To determine the baseline quality and capability of completely automatic transformation of specimen labels, a subset of the test dataset that contains each different type and quality of specimen label will go through one or more machine processes to transform the printed data to digitized form through Optical Character Recognition (OCR). The result will be a set of OCRed labels to be used in the assessment.
- Work Area 2.2. Metrics and Assessment of Machine-Processed Label Data (Months 2-5)
  The project team will develop procedures, criteria, and metrics for assessing the results of the machine-processed label data. The team will analyze the results and determine the quality compared to the criteria and metrics. In addition, this assessment will indicate user tasks that may be necessary for correcting and enhancing different types and quality of labels. The results of this work area will inform decisions about the human processes and user interfaces needed in the workflow.

Phase 3. Designing and Developing Human Process Interfaces and Their Assessment

- Work Area 3.1. Programmatic Design and Development of Interfaces (Months 5-20)
  Acting on the results of Work Area 2.2 and our preliminary workflow model and candidate processes and interfaces, we will specify the requirements for each interface, and then design and develop user interfaces to meet the requirements. The requirements will reflect the planned user tasks that should be supported by the interface and underlying functionality. Rapid prototyping and iterative development will be used; we anticipate involving the project Advisory Group in participatory design of the prototypes.
- Work Area 3.2. Usability of Interfaces (Months 5-20)
  We will develop test plans for assessing the usability of each of the prototype interfaces. Feedback from the testing will inform revisions to the prototypes. At least three types of usability assessments will be used: expert usability assessment where project team experts will systematically examine the prototype for functionality, logic, and presentation; heuristic usability assessment where project team members and volunteers will carry out tasks expected to be supported by the interface; and user testing by likely users of the interfaces. Project research assistants, BRIT staff, and dedicated volunteers from the BRIT Herbarium (who have had some exposure to concepts of specimen label data) will test interfaces and provide feedback.

Phase 4. Integrating Interfaces in Workflow and Assessment of Workflow

- Work Area 4.1. Integration of Interfaces into Workflow (Months 10-20)
  The separate interfaces from Work Area 3.1 will be incorporated into the overall workflow for transforming, correcting, and enhancing label data. As prototypes reach a relatively stable implementation, they will be brought into the workflow framework to be tested.
- Work Area 4.2. Assessment of Integrated Workflow (Months 10-20)
  We will develop test plans for systematic assessment of the evolving and final versions of the workflow, including criteria related to effectiveness, efficiency, and usability. Similar types of usability assessment listed in Work Area 3.2 will be used. Results of the assessments will inform continuous improvement and lead to a final stable version of the workflow. The testbed framework will be

improved as needed to support the workflow and interfaces and any quality assessment methods necessary.

- Work Area 4.3. Metadata Quality Assessment (Months 20-22)
  We will build on the assessment criteria in Work Area 2.2 for assessing the metadata quality of the transformed, corrected, and enhanced label data resulting from the workflow. We will draw on the literature of metadata quality assessment (e.g., Hillman and Bruce, 2004) and will establish appropriate quality measures, quality criteria, and compliance indicators.

Phase 5. Project Evaluation

Work Area 5.1. Project Evaluation (Months 22-24)

This work area addresses the overall evaluation of the project. See next section for a description of the evaluation plan.

Evaluation Plan

We anticipate the results of this research will contribute substantially to ongoing efforts in processing specimen label data. The testbed provides a critical method for deepening the natural history communities' understanding of potential solutions, and thus catalyzing changes in skills, attitudes, knowledge, and behavior. Project evaluation is in two forms, and each will be guided by evaluation plans developed by the project team:

- Evaluating the success measured against project goals and objectives: A tangible measure of the project's success is the extent to which it achieves stated objectives and answers the research questions. One measure of success is the completion of all work and deliverables for each work area. Testing is a large component of the research, and the development and implementation of rigorous methods, criteria, and metrics will be the foundation for reliable and valid results. An assessment focusing methodological rigor and reliability of the research findings is appropriate for this project.
- Evaluating outcomes: The outcomes and impact of research projects often occur well after the end of the project, once research findings are presented and published. Our dissemination of ongoing work at conferences, in papers, and articles will create the groundwork for reaching key stakeholders. The project website will provide all project documents, and regular examination of website traffic can be an indicator of the extent to which information is being accessed, the first step in changes in knowledge about new processes for label data transformation. Deepening the communities' knowledge may be manifested in longer-term changes in skills, attitudes, and behavior, as well as in demonstrated interest in collaborating on future projects based on this preliminary research.

Project Resources: Budget, Personnel, and Management

The proposed research project is a collaboration of two institutions: Texas Center for Digital Knowledge (TxCDK) at UNT and the Botanical Research Institute of Texas (BRIT). This project will facilitate a synergistic collaboration between the two institutions. Key personnel have expertise spanning the fields of botany, museum collections management, digital libraries, creation and management of digital data and metadata resources, and user interface and web service design. Both institutions are supportive of collaborative research and are providing substantial cost-sharing of $192,867.00 (20%) for the Principal Investigators (PIs) and BRIT staff time, BRIT volunteer time, support for one research assistant, and hardware purchases.

Budget

We are requesting $757,990 from IMLS to cover the costs (direct and indirect) for this research project. Funding from IMLS will be used primarily for:

- UNT graduate research assistants, including tuition stipends, under the supervision of PI Moen
- Partial salaries for the PIs and senior personnel
- Developers, a herbarium data specialist, and partial salary for a volunteer coordinator
- Travel to team meetings, professional and scholarly conferences, etc.

The Budget Justification details anticipated expenses. Funding for project staff is the largest cost category.

Personnel

The project staff will consist of three Principal Investigators: William E. Moen, Ph.D. TxCDK, UNT; Amanda K. Neill, BRIT Herbarium; Jason H. Best, Department of Research, BRIT

Moen will devote approximately 13% of his faculty time during the 24-month project.. Each year of the project, Neill will commit approximately 25% of her time, and Best will commit approximately 75% of his time. These three PIs constitute the Project Management Team, and they will manage project resources and staff at their respective institutions, with support from their institutional staff. All appropriate institutional administrative procedures will be followed related to staffing, payment of salaries, travel, and other aspects of the project where expenses will be incurred.

Moen is an Associate Professor in the School of Library and Information Sciences (SLIS) and Director of TxCDK. Dr. Moen has managed a number of large research and development projects, including two IMLS National Leadership Grants, and is currently PI for a project funded by the Texas Higher Education Coordinating Board to design and develop a learning object repository for the Board's statewide Texas Course Redesign initiative. His research focuses on metadata standards, development, assessment, and use, in addition to application development, testing, and usability. For the proposed project, Dr. Moen will lead the overall management of the project, will be involved in implementation of the machine-processing phase (Work Areas 2.1 & 2.2), and will lead the testing and evaluation of system interfaces and features (Work Area 3.1 & 3.2) including quality assessment of the label data (i.e., metadata) resulting from the workflow processes.

Amanda Neill is the Director of the Herbarium at BRIT, and Co-Director of BRIT's Andes to Amazon Biodiversity Program (previously funded by the Gordon and Betty Moore Foundation; currently funded by the U.S. National Science Foundation). She has substantial experience in training and testing staff and volunteers in both digital and nondigital tasks relating to herbarium organization, management, and databasing. She will lead the selection of the test datasets of already-digitized specimens from the BRIT Herbarium (Work Area 1.3), will oversee the assessment of success of machine and human processing work areas (Phase 2 and Work Area 4.2), and will inform the development of interfaces (Work Areas 3.1 & 3.2).

Jason Best is the IT Manager at BRIT. He was a primary creator of the Atrium Biodiversity Information System (http://www.atrium-biodiversity.org; see Supporting Document #2, developed at BRIT with funding from the Gordon and Betty Moore Foundation, and continues to oversee its development and programming. He has experience training users of web-based applications, designing and testing user interfaces, and communicating biologists' and collection managers' requirements to programmers on his team. He will lead the development and implementation of programming throughout the project, and will direct the integration of interfaces into workflow (Work Area 4.1).

Three graduate research assistants will be employed to assist in the development and testing of components and interfaces for the workflow. They will assist in usability testing and metadata quality assessment. The research assistants will develop new knowledge and skills related to research, technology, and standards and the application of these to museum specimen data issues. They will also benefit by gaining experience in real-world collaboration with UNT faculty members and BRIT staff.

Management Plan

UNT will serve as the lead institution and Dr. Moen will be responsible for the overall management of the project, including project planning, fiscal oversight, project website development, direction of graduate student research assistants, and all required reports. With support from the SLIS staff he will administer the project resources at UNT in compliance with the University's policy and procedures. PIs at BRIT will be responsible for creation of the testbed, project design implementation, and day-to-day management of BRIT staff and contract programmers. They will work with the UNT graduate research assistants to test, refine, and assess the workflow. The project team will collaborate and coordinate activities using the methods established in Startup Phase in addition to phone conferences and face-to-face meetings.

The Project Design and Evaluation Plan section presents the project's research questions, strategic work phases, and work areas. Upon award of the grant, the team will develop a detailed management plan for oversight and monitoring to guide all project activities and ensure successful and timely completion of the research. To provide external oversight of the project, we will establish a Project Advisory Group (AG) of 7-10 members, drawn from selected leaders and representatives of stakeholder groups, including botanists, collections managers, bioinformaticians, and digital library scientists. The following have made commitments to serve on the AG: Paul Berry, Director of the Herbarium, University of Michigan; Stan Blum, TDWG and California Academy of Sciences; Chris Freeland, Director of Bioinformatics, Missouri Botanical Garden; Jane Greenberg, Metadata Research Center, University of North Carolina, Chapel Hill; Mark Phillips, Manager, Digital Projects Unit, UNT Libraries; Martin Terry, Curator, Powell Herbarium, Sul Ross State University; and Sula Vanderplank, Herbarium Collections Manager, Rancho Santa Ana Botanic Garden. The AG will provide guidance to the Project Management Team by reviewing project plans, providing feedback on key issues related to the project, and reviewing project deliverables, as well as helping to disseminate information about this project.

Dissemination

Our research results will be shared with several key audiences interested in the processes and technologies for acquiring and managing data associated with museum specimens. These key audiences include managers of herbaria and other natural history collections in the Society for the Preservation of Natural History Collections and the Natural Science Collections Alliance; botanists and other biologists at the annual national Botany Conference; the biodiversity standards community (Taxonomic Database Working Group/Biodiversity Information Standards); and digital content and services managers, and the broader information science community at the Joint Conference on Digital Libraries, Annual Conference of the American Society for Information Science and Technology, and European Conference on Research and Advanced Technology for Digital Libraries. We plan to submit demonstrations and papers to be considered at conferences associated with these audiences. In addition, articles derived from the research activities will be submitted to a wide range of peer-reviewed scholarly and professional publications including: *Journal of the American Society for Information Science; International Journal on Digital Libraries; Bioinformatics*; and the *Journal of Intelligent Information Systems*.

The project will establish a website, hosted by TxCDK, for all documentation pertaining to the research activities including this proposal and progress reports to IMLS, detailed research plan and methodology, results, demonstration of user interface functions, and technical documentation for testbed components. The website will be maintained for at least three years after the end of the grant, and appropriate project documents will be placed in relevant digital repositories for future access. To broaden the reach of this project, source code developed to build the testbed (along with the technical documentation) will be made available to the programming communities by using an SVN repository such as Sourceforge or Google Code.

Sustainability

We will make project deliverables broadly available through the project website, and we will deposit selected project documents in appropriate repositories to ensure long-term access to project deliverables.

Maintenance of the testbed will sustain what was initiated and achieved in the proposed project. BRIT will continue to use and improve the testbed beyond the final deliverables of this proposal. The proposed research will provide the basis for future development of a much-needed tool to improve access to biodiversity data. It is our hope that the research can be carried forward into the development of a production-level open-source stand-alone tool or a component that could be integrated with Atrium, HERBIS, BioGeomancer (http://www.biogeomancer.org), and other web-based services in the future to aggregate expertise, increase efficiency, and reduce the duplication of effort (Atkins et al. 2003). The proposed project should be a complement and a catalyst to other ongoing efforts such as these in the information sciences for natural history collections.

References

Atkins, D.E. et al. 2003. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. (http://www.nsf.gov/od/oci/reports/toc.jsp)

Beaman, R.S., N. Cellinese, P.B. Heidorn, Y. Guo, A.M. Green, & B. Thiers. 2006. Abstract: HERBIS: Integrating digital imaging and label data capture for herbaria. 2006 Botany Conference. (http://www.2006.botanyconference.org/engine/search/index.php?func=detail&aid=402)

Daniel J. Cohen, D.J. and R. Rosenzweig, 2006. Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. Philadelphia: University of Pennsylvania Press. (http://chnm.gmu.edu/digitalhistory)

Funk, V. A. 2002. The Importance of Herbaria. *Plant Science Bulletin* 49(3): 94-95. (http://www.mnh.si.edu/biodiversity/bdg/Funk2003-Importance.pdf)

Funk, V.A. and N. Morin. 2000. A survey of the herbaria of the Southeast United States. *SIDA, Bot. Misc.* 18: 35-52. (http://www.mnh.si.edu/biodiversity/bdg/Funk&Morin2000_bestcopy.pdf)

Heidorn, P.B., Q.W. Yin, R.S. Beaman, and N. Cellinese. 2007. Abstract: Learning by example: Machine learning and herbarium label digitization. 2007 Botany Conference. (http://www.2007.botanyconference.org/engine/search/index.php?func=detail&aid=1202)

Heritage Preservation. 2005. A Public Trust at Risk: The Heritage Health Index Report on the State of America's Collections. (http://www.heritagehealthindex.org)

Hillman, D.I., and T.R. Bruce. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In: Hillman, D.I., and E.L. Westbrooks, (Eds). Metadata in Practice, American Library Association, Chicago, IL.

Holmgren, P.K., N.H. Holmgren and L.C. Barnett. 1990. Index herbariorum. Part I: The herbaria of the world. 8th edition. New York Botanical Garden. 693 pp.

Morris. P.J., 2005. Relational database design and implementation for Biodiversity Informatics. *Phyloinformatics* 7:1-63. (http://www.athro.com/general/Phyloinformatics_7_85x11.pdf)

Morris, P.J. and J.A. Macklin, 2006. Tools, techniques, and code for supporting image databases of natural history collections materials. *Collections Forum* 21:203-222.

National Science Board. 2005. Long-lived Digital Data Collections: Enabling research and education in the 21st century. NSF. (http://www.nsf.gov/pubs/2005/nsb0540)

Rabeler, R. K., and J. A. Macklin. 2006. Herbarium networks: towards creating a 'toolkit' to advance specimen data capture. *Collections Forum* 21: 223-231.

Schedule of Completion

This schedule of completion shows the project components and timeline for this 24-month research project. The projected start date is December 1, 2008; completion date is August 31, 2010. Pre-project activities will commence upon award of grant (anticipated as mid-September 2008). Pre-project activities include recruitment of graduate students for the project, securing space for project, etc. The following table provides additional detail to support the **Project Design and Evaluation** and **Project Resources: Budget, Personnel, and Management** sections of the proposal narrative.

Total direct costs requested from IMLS are $757,990. The following tables indicate how these funds will be allocated across the project phases. The method of computation for each activity was to take average monthly costs as reflected in the budget. The monthly costs were allocated as a percentage to each of the activities occurring in a month. A total cost for each activity is based on number of months each activity occurs. The costs reported are direct costs supported by IMLS funds; indirect costs and cost sharing amounts are not included in the amounts reported for the project phases.

Project Phases	Key Activities	Duration	Cost
Work Area 0: Project Startup	• Detailed project planning • Selection of graduate research assistants • Plan for website presence • Establish project advisory group • Set up communication internally among project staff and with advisory group • Preliminary technology assessment for testbed • Develop preliminary evaluation plan	2 months	**$0.00**
Phase 1. Designing the Testbed, the Workflow, and Preparing the Test Datas	• Design testbed • Design interaction between services • Design workflow framework • Implement workflow framework for machine processing • Assess specimen type dataset for extent of being representative of specimen label variation • If necessary, digitize small number of additional specimens to ensure representative test dataset • Finalize test dataset	4 months	**$52,566**
Phase 2. Assessing Quality of Machine Processing of Test Dataset	• Carry out machine-only processing of label data • Develop criteria and metrics for assessing quality of machine-only transformed label data • Assess the quality of transformed label data • Identify human process and interfaces to add to workflow	3 months	**$54,208**
Phase 3. Designing and Developing Human Process Interfaces and Their Assessment	• Specification of requirements for user interfaces • Participatory design of interfaces • Rapid prototyping of interfaces • Develop test plan for assessing interfaces • Implement assessments of interfaces including usability testing	15 months	**$269,521**
Phase 4. Integrating Interfaces in Workflow and Assessment of Workflow	• Integrate separate interfaces after assessment • Develop test plan for assessing workflow after interface integration • Implement assessments of workflow • Develop metadata quality assessment plan • Implement metadata quality assessments	13 months	**$203,927**
Phase 5. Project Evaluation and Completion	• Evaluate project against goals and objectives, and answer research questions • Evaluate project outcomes based on developed outcomes evaluation • Complete all documentation and final report	3 months	**$75,239**
Total Direct Costs from IMLS			**$655,461**

This project requires a parallel effort in multiple areas. Duration for some of the activities is the entire project (e.g., project management), while others are periodic (e.g., project evaluation). The following tables summarize duration of all work areas per project year

Year 1: December 1, 2008 – August 31, 2009

Work Area	Month		1	2	3	4	5	6	7	8	9
	Oct 2008	Nov 2008	Dec 2008	Jan 2009	Feb 2009	Mar 2009	Apr 2009	May 2009	Jun 2009	Jul 2009	Aug 2009
Work Area 0. Project Startup											
Work Area 1.1. The Testbed and Services (Months 1-4)											
Work Area 1.2. The Initial Workflow (Months 1-3)											
Work Area 1.3. The Test Dataset (Months 1-3)											
Work Area 2.1. Machine Processing of Test Dataset (Months 2-4)											
Work Area 2.2. Metrics and Assessment of Machine-Processed Label Data (Months 2-5)											
Work Area 3.1. Programmatic Design and Development of Interfaces (Months 5-20)											
Work Area 3.2. Usability of Interfaces (Months 5-20)											
Work Area 4.1. Integration of Interfaces into Workflow (Months 10-20)											
Work Area 4.2. Assessment of Integrated Workflow (Months 10-20)											
Work Area 4.3. Metadata Quality Assessment (Months 20-22)											
Work Area 5.1. Project Evaluation (Months 22-24)											

Year 2: September 1, 2009 – August 31, 2010

Month	10	11	12	13	14	15	16	17	18	19	20	21
Work Area	**Sep 2009**	**Oct 2009**	**Nov 2009**	**Dec 2009**	**Jan 2010**	**Feb 2010**	**Mar 2010**	**Apr 2010**	**May 2010**	**Jun 2010**	**Jul 2010**	**Aug 2010**
Work Area 0. Project Startup												
Work Area 1.1. The Testbed and Services (Months 1-4)												
Work Area 1.2. The Initial Workflow (Months 1-3)												
Work Area 1.3. The Test Dataset (Months 1-3)												
Work Area 2.1. Machine Processing of Test Dataset (Months 2-4)												
Work Area 2.2. Metrics and Assessment of Machine-Processed Label Data (Months 3-5)												
Work Area 3.1. Programmatic Design and Development of Interfaces (Months 5-20)												
Work Area 3.2. Usability of Interfaces (Months 5-20)												
Work Area 4.1. Integration of Interfaces into Workflow (Months 10-20)												
Work Area 4.2. Assessment of Integrated Workflow (Months 10-20)												
Work Area 4.3. Metadata Quality Assessment (Months 20-22)												
Work Area 5.1. Project Evaluation (Months 22-24)												

Year 3: September 1, 2010 – November 31, 2010

Month	22	23	24
Work Area	**Sep 2010**	**Oct 2010**	**Nov 2010**
Work Area 0. Project Startup			
Work Area 1.1. The Testbed and Services (Months 1-4)			
Work Area 1.2. The Initial Workflow (Months 1-3)			
Work Area 1.3. The Test Dataset (Months 1-3)			
Work Area 2.1. Machine Processing of Test Dataset (Months 2-4)			
Work Area 2.2. Metrics and Assessment of Machine-Processed Label Data (Months 3-5)			
Work Area 3.1. Programmatic Design and Development of Interfaces (Months 5-18)			
Work Area 3.2. Usability of Interfaces (Months 5-18)			
Work Area 4.1. Integration of Interfaces into Workflow (Months 10-18)			
Work Area 4.2. Assessment of Integrated Workflow (Months 10-20)			
Work Area 4.3. Metadata Quality Assessment (Months 20-22)			
Work Area 5.1. Project Evaluation (Months 22-24)			