September, 2015

Big things have small beginnings: Curating a large natural history collection - processes and lessons learned

Stacey Knight-Davis, *Eastern Illinois University*
Todd Bruns, *Eastern Illinois University*
Gordon Tucker, *Eastern Illinois University*

JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Big Things Have Small Beginnings: Curating a Large Natural History Collection—Processes and Lessons Learned

Stacey Knight-Davis, Todd Bruns, Gordon C. Tucker

Knight-Davis, S., Bruns, T., Tucker, G. C. (2015).Big Things Have Small Beginnings: Curating a Large Natural History Collection – Processes and Lessons Learned. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1240. http://dx.doi.org/10.7710/2162-3309.1240

ISSN 2162-3309 10.7710/2162-3309.1240

PRACTICE

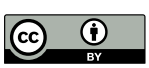
Big Things Have Small Beginnings: Curating a Large Natural History Collection–Processes and Lessons Learned

Stacey Knight-Davis
Head, Library Technology Services, Eastern Illinois University
Todd Bruns
Institutional Repository Librarian, Eastern Illinois University
Gordon C. Tucker
Professor, Department of Biological Sciences, Eastern Illinois University

INTRODUCTION Digitization of natural history collections is underway in earnest around the world and presented via platforms such as JSTOR Plants. Few natural history digital collections of specimens exist in academic institutional repositories, in spite of the fact that many universities have repositories and also hold extensive natural history collections. At Eastern Illinois University, a mid-sized public university, librarians worked with the Biological Sciences department to develop the means to digitize the 80,000 specimens of the Stover-Ebinger Herbarium collection. **DESCRIPTION OF THE PROJECT** Setting up digitization of the herbarium collection required meeting with experts associated with important projects in the field, such as Symbiota, and with acquiring the correct digitization equipment. Data management techniques had to be developed to move metadata from an Access database to Symbiota and to the institutional repository platform. These were informative steps to be taken and will enable easier development of future natural history collections. **NEXT STEPS** Having procured the correct equipment and expert guidance, the library is ready for move forward with digitization of this large collection. The existing 16,000 records in the repository will have images added to them, databasing and imaging will proceed for the remaining 64,000 specimens, and we will be exploring the impact of these specimen records in the "Cited by" notation in Google Scholar, as well as adding specimen field notes to enhance the collection.

Received: 03/01/2015 Accepted: 06/12/2015

Correspondence: Stacey Knight-Davis, Booth Library, 600 Lincoln Avenue, Charleston, IL 61920, slknight@eiu.edu



INTRODUCTION

> "Big things have small beginnings, sir."
> Claude Rains in *Lawrence of Arabia*

By the fall of 2013 our institutional repository (IR), The Keep (http://thekeep.eiu.edu), had been in operation for two years. A rapidly growing IR, The Keep had already reached 27,000 records, largely through aggressive digitization of archives collections, ingestion of already existing digital content (master's theses and archival photos), and the involvement of approximately 85 "early adopter" faculty members.[1] Among the early adopters, the department of Biological Sciences was a particular standout, with participation reaching 88% of the department's faculty. Familiarity with The Keep as a potential resource was therefore high among the Biological Sciences faculty and chair.

The Biological Sciences department's Stover-Ebinger Herbarium includes approximately 80,000 specimens. The Herbarium was founded in 1899 and is currently curated by Dr. Gordon Tucker. Each specimen in the collection is a portion of a plant that has been carefully preserved by drying and then mounted to a heavy sheet of paper. A label detailing the location, collection site conditions, collector's name, and other details is attached to each sheet. The label also generally contains a number to reference the specimen back to notes in the collector's field journal. The Stover-Ebinger Herbarium collection is used for teaching and includes locally collected plants, as well as global specimens provided by other herbaria around the world.

Students learning in a well-curated herbarium are able to touch and see plants from countries to which they may never travel. These collections are built over generations, with some specimens in the Stover-Ebinger Herbarium dating from the early 19th century. Among the specimens are items collected by botanists renown in the field, including Edgar Nelson Transeau, Mary Agnes Chase, and Douglas Barton Osbourne Savile. Until recently, this valuable collection could only be searched on site.

In the fall of 2013, the chair of Biological Sciences asked the IR librarian about digitizing the herbarium collection and including it in The Keep. A meeting between the IR librarian and Herbarium Curator Dr. Tucker thus began a project that would represent the maturing of The Keep into a substantial repository, involve both the IR librarian and the

[1] The Keep statistics as of fall 2013 are available in this report: http://works.bepress.com/todd_bruns/36/

Head of Library Technology Services, and require steep learning curves in a number of areas including equipment procurement, metadata schema, data manipulation, and cross-platform communication. By opening up the collection for discovery, scholars around the world would see what is available and potentially request the loan of the physical sheets for further study or genetic analysis. This paper describes the process of digitizing this valuable content to make it available for researchers and students around the world.

LITERATURE REVIEW

While there is a growing movement to digitize herbaria collections, there is not yet a national single portal of all herbaria collections in the United States (Barkworth & Murrell, 2012). Global initiatives such as JSTOR's Global Plants[2] and the Global Biodiversity Information Facility[3] digitize and provide biodiversity data. JSTOR's Global Plants originally began as the Global Plants Initiative, funded by a grant from the Andrew Mellon Foundation, and has since grown to over two million type specimens. It is described as the largest community-contributed database of plants, available to institutions for subscription. In contrast to the subscription based model of Global Plants, the World Flora Online[4] project seeks to create an Open Access resource that includes all known plants. The project is a collaboration between the Missouri Botanical Garden, New York Botanical Garden, Royal Botanic Garden Edinburgh and the Royal Botanic Gardens, Kew. Funding from the Sloan Foundation in 2015 will make World Flora Online available as an Open Access resource through the Digital Public Library of America. This ambitious project has a target completion date of 2020 (New York Botanical Garden, 2015).

World Flora Online and Global Plants provide a catalog of plants, but they do not include all collected instances of each plant. The Global Biodiversity Information Facility (GBIF) offers a portal to combined biodiversity data for all types of organisms from many institutions. GBIF is also Open Access and contains 500 million records representing over 1.6 million species (of which, over 1.2 million are plant species).

Library participation in the digitization and presentation of natural history collections has thus far been focused on the digitization of natural history documents rather than specimens. Although libraries are increasingly heavily invested in digital collections and

[2] JSTOR's Global Plants: https://plants.jstor.org/

[3] Global Biodiversity Information Facility: http://www.gbif.org

[4] http://www.missouribotanicalgarden.org/plant-science/plant-science/world-flora-online.aspx

supporting digital scholarship, the majority of library digital collections are of content the library owns rather than collaborative efforts with other campus departments (Schmidt, 2007). A review of the literature reveals few natural history collection specimens associated with institutional repositories, in spite of the fact that most research institutions have IRs (Dubinsky, 2014), and most herbaria collections are housed at academic institutions (Barkworth & Murrell, 2012).[5] The majority of herbaria digital collections at academic institutions are either hosted on locally developed platforms or are on ContentDM, a platform option that predates repository software.[6]

The only large herbarium collection hosted in an institutional repository that could be discovered searching the literature or through a web search is Wichita State University's Herbarium[7] (Jameson & Matveyeva, 2010), built on the D-Space platform. Existing herbaria collections in repositories built on Digital Commons are, to date, small collections.[8] Repositories, however, offer some advantages over hosting collections in ContentDM, including search engine optimization,[9] connecting the content contextually with the institution as part of the institutional repository rather than a library digital collection, and pre-configured Google Maps integration for displaying the location where the specimen was collected.

There is a growing research need for access to these materials. Digitizing natural history collections like herbaria provides access to collections that are largely otherwise inaccessible. It also increases the visibility of institutional collections and reduces the need for physical

[5] It should be noted that while most herbaria collections are at institutions of higher education, the majority of specimens are in museums/botanical garden collections. The *Index Herbariorum: A Global Directory of Public Herbaria and Associated Staff* (http://sciweb.nybg.org/science2/IndexHerbariorum.asp) provides a scope of the extent of herbaria collections worldwide.

[6] Examples include Butler University Friesner Herbarium Digital Collection, http://palni.contentdm.oclc.org/cdm/landingpage/collection/herbarium4 and Rossbach Herbarium, http://cdm16111.contentdm.oclc.org/cdm/landingpage/collection/p15135coll5

[7] This collection is available in the Wichita State University institutional repository, SOAR: http://soar.wichita.edu/handle/10057/2861/browse?value=Wichita+State+University%27s+Herbarium&type=subject

[8] An example of a Digital Commons hosted herbarium collection is Utah State: http://digitalcommons.usu.edu/herb_typspec/

[9] An example of how repository software provide search engine optimization (the example given is Digital Commons) is available here: http://digitalcommons.bepress.com/cgi/viewcontent.cgi?article=1040&context=reference

handling of these specimens (Schmidt, 2007). Natural history collections are a valuable resource for the study of biodiversity informatics and changes in organism distribution and morphology (Blagoderov, Kitching, Livermore, Simonsen, & Smith, 2012; Feeley, 2012; Lister, et al., 2011; Park, 2012; Shaffer, Fisher, & Davidson, 1998), but are also not without inherent issues of potentially introducing bias into a study, via errors such as "false presences"[10] and vague or incorrect geophysical locating (Graham, Ferrier, Huettman, Moritz, & Peterson, 2004; Tingley & Beissinger, 2009). To correct for this, Tingley and Beissinger (2009) call for observational and occurrence data from field notes to be given greater value and made publically available as part of the important metadata of a record. The Smithsonian Institute's Field Book Project[11] and the Biodiversity Heritage Library,[12] both containing significant collections of digitized field notes, are important steps in this direction.

The usage of historical data for studying species decline (Shaffer, Fisher, & Davidson, 1998) has become more urgent in recent decades due to climate change (Feeley, 2012; Park, 2012; Tingley, & Beissinger, 2009). With growing evidence that there is now a sixth mass extinction of animals and plants underway (Barnosky, et al., 2011), making historical natural history collections available for researchers worldwide is an important task for institutions to undertake.

CHALLENGES OF DATA CAPTURE

An herbarium specimen is a thin, fragile three-dimensional object. Flatbed scanners, the main image capture technology used in the Booth Library Scanning Center, project light up from the bottom while the object being imaged lies face down. Flatbed scanning is completely inappropriate for herbarium specimens. Damage to specimens is very likely to occur if the specimen sheet is turned over. The specimen illustrated in Figure 1 (following page), *Eleocharis acicularis*, is composed of many fine hair-like spikelets. When face up, the spikelets are supported by the sheet. When turned over, the specimen hangs from the glue attaching it to the sheet. There is great risk of pieces falling off or glue breaking if a specimen is flipped over.

[10] A "false presence" is the misidentification of a species.

[11] The Smithsonian Institute's Field Book Project (http://www.mnh.si.edu/rc/fieldbooks) is cataloging and digitizing field books, making them available via the Institute's Collection Search Center (http://collections.si.edu/search/) and the Digital Public Library of America (http://dp.la/).

[12] The Biodiversity Heritage Library (http://www.biodiversitylibrary.org) is a consortium of natural history and botanical libraries working in concert to digitize biodiversity literature held in their collections.

Figure 1. A sample herbarium specimen, Eleocharis acicularis. This sheet still requires barcoding. *Image credit*: Gordon Tucker

Even if the specimen could be safely turned over, the image produced by a flatbed scanner contains many undesirable shadows because the light projects up on to the three dimensional specimen. The best solution is to capture the image from above. The "Herbscan" was an attempt to safely use a flatbed scanner for herbarium specimen image capture. The Herbscan was developed by the Royal Botanic Gardens, Kew, and was the required image capture system for institutions participating in the Global Plants Initiative. The Herbscan consists of a flatbed scanner turned upside down and mounted in a custom built cabinet. The

specimen is placed in the cabinet on a platform with a foam cushion that is raised up to meet the glass of the scanner. The light comes from above, but raising and lowering the platform is time consuming.

A digital camera can also capture images and allows the specimen to be lit from the top. However, the camera must be capable of acquiring a very high resolution image. The project team attempted to use a 12 megapixel camera for image capture and quickly discovered that it was not capable of producing the resolution needed for successful optical character recognition.

In addition to perfecting image capture, it was also necessary to convert the existing specimen metadata for 16,000 specimens contained in a Microsoft Access database so that it could be imported to the IR. Determining how to best present the metadata for individual specimens through The Keep, as well as deciding on what formats in which to provide the full dataset, was especially challenging. Extensive searching in library science and biological sciences literature for a standard set of metadata elements used to describe herbarium specimens did not provide a solution. We then sought expert assistance from Susan Braxton, Head Institute Librarian at the Prairie Research Institute, University of Illinois at Urbana-Champaign. She suggested iDigBio (https://www.idigbio.org/) which provided the needed answers to data and imaging problems.

iDigBio is the short name for Integrated Digitized Biocollections, the National Resources for Advancing Biodiversity Collections funded by the National Science Foundation. The iDigBio website provides excellent guidance on best practices for digitizing various types of natural history collections. Equipment specifications, required resolution, file formats, sample workflows and more can be found here.[13] While the Scanning Center had a long record of following local, state, and consortial standards for digitizing archival photos and text documents, the newly discovered standards for producing biological images were much more demanding.

Literature in iDigBio centered on two options for image capture. The Herbscan was developed for the Global Plants Initiative[14] (later JSTOR's Global Plants), and was provided to institutions receiving funding from the Andrew W. Mellon Foundation to digitize "type specimens," the specimen that a species description is based on. Alternatively, a camera

[13] https://www.idigbio.org/wiki/index.php/Digitization_Resources

[14] More information on this collaboration can be found at http://about.jstor.org/news/global-plants-initiative-and-jstor-release-online-database-study-plants

capable of capturing at least 20 megapixels mounted on a copy stand equipped with special lenses and filters and a non-directional light source would also produce images of the right resolution, size, and color depth. As outlined by Tulig, Tarnowsky, Bevans, Kirchgessner, and Thiers (2012), each image must also include a ruler and a color target.

With the technical demands now clear, the question of data format was addressed. In addition to iDigBio, Susan Braxton also suggested further investigation of the Darwin Core family of standards. After learning more about Darwin Core, it quickly became clear that the flaw in the earlier searches for a metadata standard had failed due to key words "herbarium" and "plants." Darwin Core goes far beyond the plant kingdom. Currently at version 1.4, the Darwin Core group of standards originated in 1998 and can be used to describe any type of biological specimen.[15] A key component is the standard set of terms to describe taxonomy, geography, specimen type, and other critical data elements. We now had the right tool to share the Stover-Ebinger data.

Further investigation into equipment and best workflow practices led the Head of Library Technology, the IR librarian, and the staff member overseeing the digitization process to meet with Dr. Andrew Miller, Director of the Herbarium/Fungarium and Fungi Curator, of the Illinois Natural History Survey (INHS) at the University of Illinois at Urbana-Champaign. Dr. Miller demonstrated two pieces of equipment: the Herbscan and a camera in a copy stand with a light-box. The light-box is a key element to making the camera setup work—it produces equal levels of light on the specimen, reducing shadows and bringing out the necessary detail of the plant. In the experience of the staff at INHS, the camera setup allows for much faster processing and less potential for specimen damage.

The INHS staff workflow involved capturing the digital image of the specimen, entering basic "skeletal record" information from the specimen label into an Excel spreadsheet and uploading that spreadsheet information every night to the Mycology Collections portal, an online search system for North American fungi powered by Symbiota.[16] Dr. Miller emphasized the importance of digitizing natural history collections and adding them to networked biodiversity information systems. Researchers can then combine the data from several herbaria collections to study changes in population size and range, reactions to climate, and other indicators of change (Graham, Ferrier, Huettman, Moritz, & Peterson, 2004; Lister, et al., 2011; Shaffer, Fisher, & Davidson, 1998).

[15] Darwin Core: http://rs.tdwg.org/dwc/

[16] Symbiota is an open source software project aimed at building a collection of web tools to aid biologists in the building of virtual collections of flora and fauna: http://symbiota.org/docs/

Armed with this new information, the project team worked to procure the necessary equipment. Since the Herbscan system was neither commercially available nor highly recommended by the INHS staff due to slower speed in use, a camera setup was chosen.

Following the equipment recommendations from iDigBio, the following items were purchased: Canon EOS 6D 20.2 MP CMOS Digital SLR camera, Canon EF 50mm f/2.5 Compact Macro Lens, Canon Life Size Converter EF Macro Lens, remote trigger, and AC power supply. The MK Photo-eBox Bio™ (New York Botanical Garden Modified), the same system used by the Illinois Natural History Survey, was selected as the lighting system. The equipment and camera settings are similar to those utilized at the William and Lynda Steere Herbarium at the New York Botanical Garden (Tulig, Tarnowsky, Bevans, Kirchgessner, & Thiers, 2012), and use Michael Bevans' recommendations for digitizing plant specimens.[17] With the equipment order in place, another major challenge was undertaken: Understanding and managing the data.

WRESTLING WITH DATA: COLLABORATIONS ACROSS CAMPUS, INSTITUTIONS, AND PLATFORMS

Digitizing herbaria is happening worldwide and involves a lot of different players (Barkworth & Murrell, 2012). One task before the project team was to understand where the data existed. It was already known that the Stover-Ebinger collection itself, the actual specimens, comprised 80,000 items. The metadata of over 16,000 of these had been entered into the aforementioned Access database maintained by Dr. Tucker. These 16,000 records presented an opportunity for an initial batch upload into the repository. The task at hand was to get the data from Access into Darwin Core. Remembering Andrew Miller's demonstration of the upload procedures for the Mycology Collections portal on the Symbiota platform, the Head of Library Technology began searching to see if Symbiota could provide a solution for us as well.

Symbiota is an open source and modular software platform. Its designers offer the following description:

> Symbiota is a leading Open Source platform in North America for mobilizing, integrating, and using specimen- and observation-based occurrence records and derivative products. (Gries, C., Gilbert, E., & Franz, N., 2014)

[17] Bevans' *So, you want to digitize plant specimens* is available on the iDigBio site: https://www.idigbio.org/wiki/images/8/82/Imaging_Manual.pdf

One of the goals of the Symbiota project is to allow groups of institutions to pool their resources to provide access to their shared data. Symbiota relies on other open source software for databasing and processing, so there is no software cost. Using the Symbiota platform allows these consortia to build a platform that meets their needs with a low cost of development. Consortia are generally regionally focused, but there are also groups, such as the Mycology Portal, that focus on a particular group of organisms. These geographically or thematically structured collaborations are known as "portals." Once data is ingested through a Symbiota portal, it can be downloaded back out of the system as Darwin Core formatted data.

Prior to the library's involvement with the herbarium, Mary Barkworth, Director, Intermountain Herbarium, Utah State University, had come to the herbarium to provide information and training on the Consortium of Northern Great Plains Herbaria Symbiota portal. After the training, all new specimen records for the Stover-Ebinger Herbarium were added on the portal. Over 500 records were keyed into Symbiota.

Initial contact with the Symbiota team was made through Dr. Edward Gilbert, Head of the Global Institute of Sustainability at Arizona State University. Dr. Gilbert brought in Ben Brandt, Scientific Software Engineer at Arizona State University and a botanist and programmer with the Symbiota project. Brandt detailed how to manipulate the existing data for a smooth import to Symbiota. While the data contained in the Access database was a remarkably good match for the Symbiota fields, there were a few differences that had to be massaged out before import. For example, the Access database held the determiner name and year of determination in the same field. To get the name and year separated, the data was imported to Excel where a formula was applied to copy the year into a separate field. Also, legal descriptions, a geographical description system based on segments of a township, was a field in the Access database that had no directly corresponding field in Symbiota. Since this important geographical data should not be lost, legal descriptions were concatenated to the Locality field, again using an Excel formula. The Access database contained no separate field for elevation, but elevation details were included as a part of various notes and locality fields. The Head of Library Technology scanned the database manually to find elevation data and then copied it to a new column that would map to the verbatim elevation field. Associated collectors were split off from the primary collectors, and notes on plants growing nearby were copied to the Associated Taxa field. The Head of Library Technology collaborated with the Herbarium Curator to get a list of the active projects that are preserved in Symbiota for label printing. To populate the scientific name field, genus/species/infraspecific[18] epithets were concatenated in one field. This was again done with an Excel formula.

[18] "Infraspecific" in botany refers to any taxon below the level of species.

In addition to structuring the data as Darwin Core, the Symbiota system ingestion also standardized the scientific name of each plant. The accepted name of an organism can change over time. To ensure that all specimens of the same plant have the same name, data in the scientific name field is mapped to the Symbiota taxonomic thesaurus. The Kingdom, Phylum, Class, Order, Family, and Scientific Name fields are automatically filled based on the thesaurus mapping. The scientific name authorship field, listing the botanist who first published the name, is also standardized when the data is ingested.

After all new columns were set up in Excel to map to their Symbiota equivalent, the data was exported as a Unicode[19] text file to preserve the diacritics and special characters. This tab-delimited file was then further manipulated in Notepad++ to convert it to a comma-delimited Unicode file. Brandt then imported the CSV file and appended it to the 567 records already entered by Dr. Tucker. Once data is ingested into one Symbiota portal, it can easily be included in other appropriate portals. Brandt copied the data to the Consortia of Midwest Herbaria portal, the Consortium of Intermountain Herbaria Collections, and other Symbiota portals.

With all records now combined, plant names standardized, and name authorship standardized, a backup of the data from Symbiota was downloaded as a CSV file. A list of threatened and endangered plants in Illinois was loaded to Access as a table. The Symbiota data was then queried against this table to identify sensitive records. Location information was deleted from these records to protect the plant from potential theft, damage, or over collecting in the wild. In the future, there is a field in Symbiota for locality security that may be used to streamline this process. The data was then exported out of Access to Excel.

The entire dataset for the Stover-Ebinger Herbarium is provided for public download as a Darwin Core Archive file.[20] Symbiota automatically generates this archive from the administrative console. The archive is a ZIP file containing comma-delimited files of the occurrence records, image data, determination history, and an xml file describing the data included in the file. Since a CSV file is such a basic file type, any user, regardless of platform, should be able to download, extract, and import the data. With standardized field names users can easily combine multiple Darwin Core datasets.

[19] It should be noted that non-Western characters are not supported by Symbiota, therefore the information must be re-entered for use in the repository.

[20] Researchers citing this dataset have available the Digital Commons default “Recommended Citation”: http://thekeep.eiu.edu/herbarium_data_files/1/

Converting the Darwin Core formatted data into individual records in Digital Commons was the next step. The Head of Library Technology and the IR librarian worked with support staff at Berkeley Electronic Press (bepress) to develop the appropriate metadata fields for specimen records in the Stover-Ebinger collection within Digital Commons. First, the new document type "Specimen" was created. To accurately describe each new Specimen document, all Darwin Core fields from Symbiota were created as added fields in Digital Commons.[21] The basic required Digital Commons fields of Title and Author now had to be filled. On consultation with Biology librarian Kirstin Duffin, the Darwin Core field recordedBy, was mapped to first author. Genus, specificepithet, and scientificNameAuthorship fields were concatenated to create the Title field. Since some sheets lack a notation for collector or collection date, bepress made these fields optional. A test site was set up on the demo server site for The Keep, allowing manipulations of metadata and format without being in actual production. Decisions were made about which fields to make visible to viewers and which to include in export labels for data harvesters. Certain repetitive fields, such as the Year, Month, and Day breakdown of the eventDate field, were not thought to be helpful to users and were hidden. Basis of Record was hidden, as all records are based on preserved specimens, and Institution code was hidden as all records are from the same institution.

Collector and associated collector names were standardized using the Harvard Name Authority.[22] Discussions of how best to present the data and images of the specimen sheets resulted in use of the Digital Commons "book template" for presentation purposes, a structure that allows for visual presentation of the specimen sheet as the "book cover," presentation of important metadata fields in the record, and the ability to add supplemental content such as additional pictures or notes.

For individual specimen records the default Digital Commons "Recommended Citation" was deemed inadequate. In consultation with Dr. Tucker, a custom citation was created that included Family, Genus, Species, Authority, Country, State, County, Location, Latitude, Longitude, Collection Date (Day, Month, Year), Collector, Field Number, and Herbarium.

In April 2015 the initial 15,722 metadata records were available in The Keep.[23] The full set of records is available in the "Specimens by Name" structure. This large collection is filtered into smaller structures: "Specimens by Collector," "Specimens by Family," and "Specimens

[21] More information about Darwin Core metadata fields is available here: http://rs.tdwg.org/dwc/terms/index.htm#theterms

[22] The Harvard Name Authority is at http://kiki.huh.harvard.edu/databases/botanist_index.html

[23] http://thekeep.eiu.edu/herbarium/

by Country." The United States segment of Specimens by Country is subdivided by state, and the Illinois structure further subdivided by county. Although in some ways the project team is still very much at the beginning of this large project, significant lessons had been learned about managing data and processing unique content for the IR (see Figure 2 for the workflow of the batch upload from the Access database to the repository).

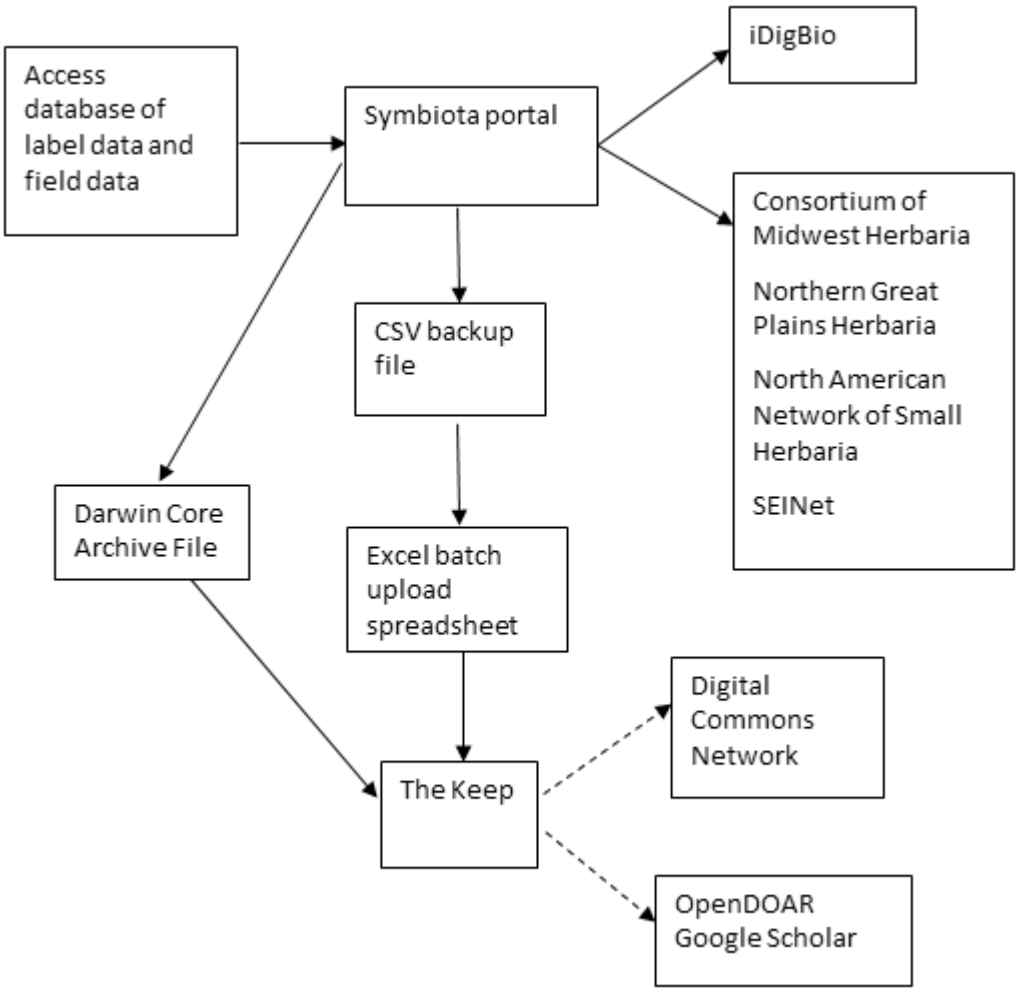


Figure 2. Data flow diagram for the initial load to Digital Commons.

LESSONS LEARNED

Tackling such a large project with content unique to the repository proved to be challenging and instructional. The project team took several important lessons from the experience:

Get the Right Equipment

The usual equipment for handling IR materials proved to be completely inadequate for this project. Our investigations revealed the necessity of acquiring a light-box and camera (Tulig, Tarnowsky, Bevans, Kirchgessner, & Thiers, 2012). Procuring the equipment proved somewhat challenging in these economic times, however we were able to secure partial funding through a campus grant program. Having the right equipment was absolutely essential for the project to happen at all, and considering the importance of the collection to the study of changes in plant populations and climate change impacts, leaving the collection largely undiscoverable was not an option.

Be Prepared to Really Work the Data

The nature of this collection required that the data be available on a number of different platforms. It wasn't enough to digitize the content and add it to the IR with a specimen name. In order for the data to really be useful, it had to be available via the larger specimen repositories as well as the IR, necessitating advanced metadata with the use of Darwin Core, and detailed manipulation of the data for export and mapping between the IR and Symbiota.

Some problems in the data are hard to spot in the raw data but are clearly apparent once live on a web platform. For example, there were variations in the spacing and punctuation of "U.S.A." When the column was quickly scanned by eye in Access, these discrepancies were not obvious. Once loaded into Symbiota, a country link was generated for each variation. This had to be standardized after the initial load to Symbiota.

Another unexpected data issue was discovered when the geolocate mapping feature was turned on for herbarium collections in Digital Commons. Data was copied directly from the decimal_latitude and decimal_longitude fields from Symbiota to populate longitude and latitude for gelolcation in the IR. Once the maps were added, it was clear that plants that should have been in Illinois were instead geolocated in China or in the southern Pacific Ocean. We quickly realized that the sign on the coordinates was not correct in these cases. To fix the issue, the specimen records were pulled back down and sorted by country and longitude. The most common problem was Western Hemisphere locations with positive longitude. These values were multiplied by -1 in Excel to correct the sign.

Moving forward, all staff working with geolocation data will be given an image illustrating the correct sign for latitude and longitude in each hemisphere.

Loading almost 16,000 records was an exercise in patience. Records were loaded in batches of 500. On an average day, two batches of 500 were uploaded. As expected, the project took a little over one month to complete. Luckily, batch revisions to fix problems like the latitude and longitude issue mentioned above finished in less than an hour. Setting up filters to divide the Specimens by Name collection into smaller sub-collections was also a time intensive process, but once the filters were set up, new material added to the parent collection automatically refreshes overnight.

Collaborate, Collaborate, Collaborate

Particularly when working with datasets outside of the field of library and information studies, it is essential to collaborate with the discipline faculty on campus as well as to reach out for collaborations with colleagues off campus. Working this data would not have been possible without the involvement of Dr. Tucker and assistance of the, Dr. Gilbert, and Ben Brandt. Their expertise in the field and their experience working with this type of data greatly informed our ability to successfully manage the dataset.

There is Always Room for Improvement

Hours after the initial upload was complete, a new file was pulled down from Symbiota to find records added to the collection in the weeks spent on the initial load. In cases where the specimen collector had also authored a thesis at EIU, links were added connecting the collector page to the thesis. Another enhancement made within days of the initial load was adding biographical information and portraits to collection pages for notable collectors. In many cases, portraits were already available via University Archives image collections hosted in the repository. Adding content to collector pages also provided an opportunity to highlight material in unexpected collections in the repository, an example being the handwritten class list containing Sargent from the 1906 EIU Senior Scrapbook, now included with the biographical material on Sargent's collector page in the Herbarium.[24]

NEXT STEPS

As of spring 2015 the project team is on the cusp of earnest digitization and production efforts on this collection. The camera and lenses have been received and are being installed

[24] http://thekeep.eiu.edu/herbarium_specimens_sargent/

and barcoded specimen sheets are waiting for digitization. With the equipment and metadata data in place, the first phase of the herbarium project in The Keep will be to connect digitized images of specimens with the existing 16,000 metadata records in the repository. The project team will analyze and evaluate this first phase of the collection prior to moving on to developing the metadata records and digital images of the remaining 64,000 specimens.[25] As daily digitizing of the specimens has not yet begun at the time of writing, we cannot offer an estimated timeframe for the completion of this project, however the project will have a dedicated computer station in the Scanning Center, allowing continual productivity.

Each specimen to be digitized has an optical barcode attached. This barcode will be searched against the existing specimen dataset to confirm that the specimen is in the system. At the point of image capture, the specimen's barcode will be scanned into an Excel spreadsheet with a handheld optical barcode scanner. We will then take a digital image, saving a digital master to our local storage array (storage considerations can be a factor in this process, as noted by Tegelberg, Mononen, and Saarenmaa (2014)). From the master, we will create a jpeg for use as a cover image in The Keep. We will also generate a PDF file. Any PDF with recognizable text will have an optical character recognition process run. PDF and jpeg files will be saved to the library's webserver. Once a substantial number of images are saved, the collection will be pulled down from Digital Commons for a batch revision. Using Access, the scanned barcode spreadsheet will be matched to the full dataset. Excel formula generated URLs for the cover image URL and full text PDF file will be inserted. The Excel sheet will then be uploaded back to Digital Commons for processing. There are multiple methods for associating images with Symbiota records. The project team will work with Symbiota support to determine which method will be the most efficient.

As we move past specimens with existing digital data, the project team will fully develop workflows for capturing specimen label data to submit to Symbiota. The current plan is to have student workers manually enter label data into appropriately named columns in Excel as part of the digitization process. The Excel file will then be converted to a csv file and uploaded to Symbiota. Once processed, the full set will be pulled back down. Newly added records will be identified by their modified date. Necessary bepress fields, cover image URLs, and full text URLs will be added using a formula. The sheet will then be uploaded for processing.

[25] The "Note from Nature" tool provides an option for accomplishing the task of digitizing large natural history collections: http://www.notesfromnature.org/#/about. See the References page for more information: Hill et al., 2012.

In addition to adding to the collection, there are other paths of development and research to peruse. We know that Google Scholar indexes the records in The Keep. We will be watching with interest to see if the "Cited by" feature works for specimens cited in other works. We will also be watching to see if there is an increase in the number specimen loan requests and inquiries about herbarium specimens.

As data enhancement, we may further investigate converting Illinois legal descriptions to latitude and longitude. Automation of this process is possible, but it is complex. However, the addition of this information adds utility for other researchers. We will further investigate linking specimen records to related master's theses and articles in The Keep, adding a depth of data richness for scholars using these materials. Initial efforts in linking specimen records to articles are visible in two articles by Gordon Tucker.[26] In keeping with the call for the importance of adding field notes (figure 6) to metadata (Tingley and Beissinger, 2009), we will also explore the potential of adding field notes as supplemental files to specimen records.

CONCLUSION

Digital scholarship is rapidly becoming the primary form of research for many scholars. Researchers and scholars need online digital access to materials, and more libraries and IRs will be called upon to make the collections and scholarship at their institutions available for the world. This is particularly important in the case of natural history collections, which remain largely hidden from view until they are digitized (Schmidt, 2007). Future IR collections are likely to hold content that thus far has been rare in repositories: natural history collections, biological specimens, 3D realia, and more. Among the calls for development in the area of digitizing herbaria are for more communication between the botanists/specialists knowledgeable about the specimens and those who best understand digitizing and presenting information (Barkworth & Murrell, 2012). Librarians' knowledge of digitizing standards, metadata, and authority control, combined with the availability of institutional repository platforms, provides a solution to this growing need.

By understanding the processes for handling the digitization and presentation of these collections, libraries will continue to be an integral part of the digital scholarship of every field. Sharing stumbles and failures, as well as successes, in the development of best practices in this area will prove pivotal to this 21st century style of librarianship.

[26] http://thekeep.eiu.edu/bio_fac/171/, http://thekeep.eiu.edu/bio_fac/170/

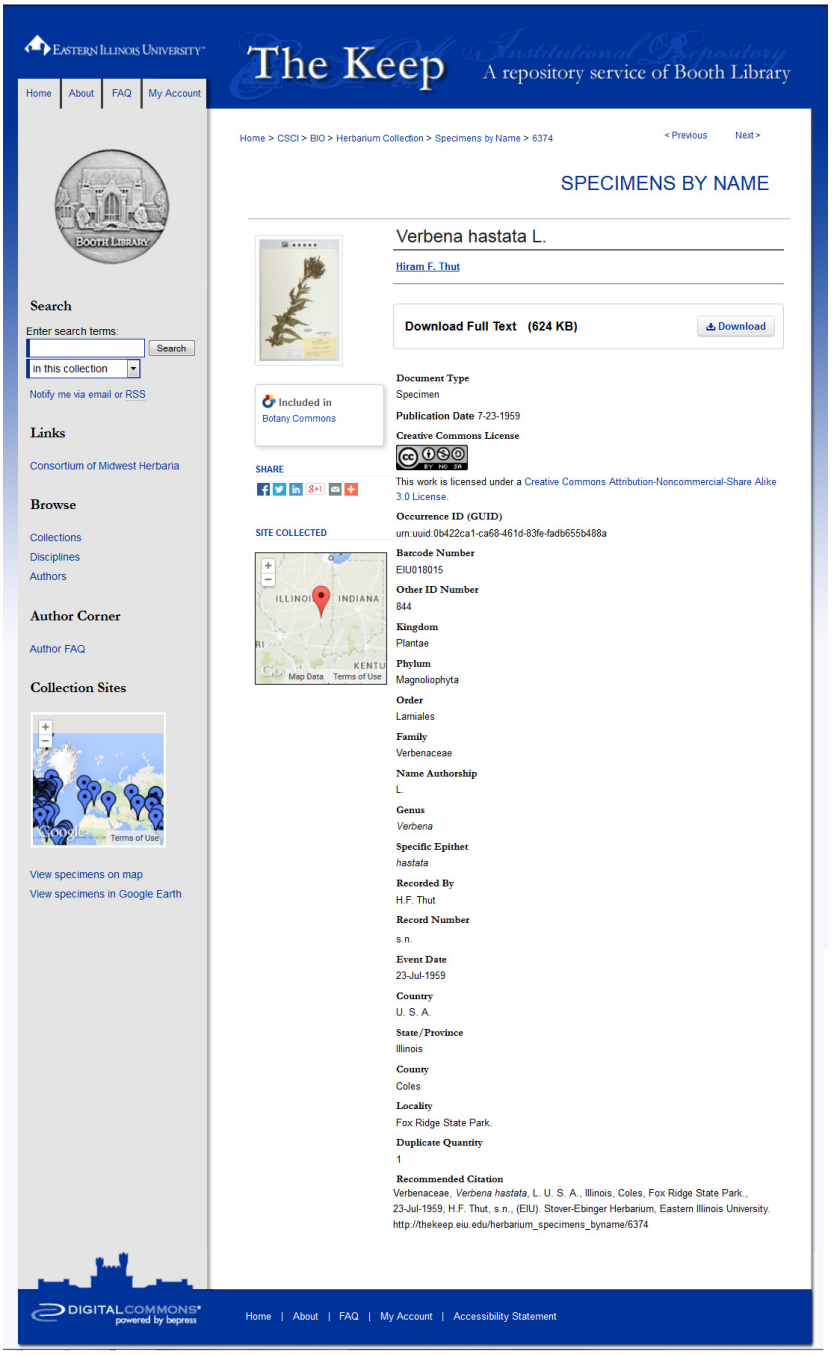


Figure 3. Screenshot of completed record.
http://thekeep.eiu.edu/herbarium_specimens_byname/6374/

ACKNOWLEDGEMENTS

This project was possible because of the invaluable assistance the authors received from several generous people. Susan Braxton, Head Institute Librarian at the Prairie Research Institute, University of Illinois at Urbana-Champaign, provided guidance on metadata standards and facilitated the meeting with INHS Herbarium staff. Andrew Miller, Director of the Herbarium/Fungarium and Fungi Curator, of the Illinois Natural History Survey (INHS), and the other wonderful staff members and researchers we met at the INHS Herbarium provided excellent guidance on digitization processes, equipment, and workflows. We greatly appreciate the time taken to explain the importance of sharing biodiversity information and advice on practical ways to make data available digitally. Edward Gilbert, Head of the Global Institute of Sustainability at Arizona State University, and Ben Brandt, Scientific Software Engineer at Arizona State University made it possible for us to have standardized plant names and data formatted correctly in Darwin Core. Special thanks to Ben for all his work on ingesting data to Symbiota. Tina Jenkins, Scanning Center, Booth Library, Eastern Illinois University was a driving force in the early month of the project. Tina identified many problems and asked all the right questions. Sarah Johnson and Kirstin Duffin offered many helpful suggested changes and editorial commentary. Equipment for this project was partially funded through a grant from the Eastern Illinois University Faculty Development Office.

REFERENCES

Barkworth, M. E., & Murrell, Z. E. (2012). The US virtual herbarium: Working with individual herbaria to build a national resource. *ZooKeys*, *209*, 55-73. http://dx.doi.org/10.3897/zookeys.209.3205

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., ... Ferrer, E. A. (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, *471*(7336), 51-57. http://dx.doi.org/10.1038/nature09678

Blagoderov, A., Kitching, I. J., Livermore, L., Simonsen, T. J., & Smith, V. S. (2012). No specimen left behind: Industrial scale digitization of natural history collections. *ZooKeys*, *209*, 133-146. http://dx.doi.org/10.3897/zookeys.209.3178

Dubinsky, E. (2014). A current snapshot of institutional repositories: Growth rate, disciplinary content, and faculty contributions. *Journal of Librarianship and Scholarly Communication*, *2*(3), 1-22, eP1167. http://dx.doi.org/10.7710/2162-3309.1167

Feeley, K. (2012). Distribution migrations, expansions, and contractions of tropical plant species as revealed in dated herbarium records. *Global Change Biology*, *18*, 1335-1341. http://dx.doi.org/10.1111/j.1365-2486.2011.02602.x

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *TRENDS in Ecology and Evolution*, *19*(9), 497-503. http://dx.doi.org/10.1016/j.tree.2004.07.006

Gries, C., Gilbert, E., & Franz, N. (2014). Symbiota – A virtual platform for creating voucher-based biodiversity information communities. *Biodiversity Data Journal*, *2*, e1114. http://dx.doi.org/10.3897/BDJ.2.e1114

Hill, A., Guralnick, R., Smith, A., Sallans, A., Gillespie, R., Denslow, M., ... Fortson, L. (2012). The notes from nature tool for unlocking biodiversity records from museum records through citizen science. *ZooKeys*, *209*, 219-233. http://dx.doi.org/10.3897/zookeys.209.3472

Jameson, M. L., & Matveyeva, S. (2010). Wichita State University's herbarium: Filling a critical gap. Presentation at the *Great Plains Plant Systemics Meeting*, October 8, 2010.

Knight-Davis, S., & Bruns, T. A. (2014). Natural history collections: Connecting with faculty and content. Presentation at *Digital Commons+ Great Lakes User Group (DCGLUG)* meeting, in Valparasio, IN, Aug. 2014. Retrieved from http://works.bepress.com/todd_bruns/45

Lister, A. M., & Climate Change Research Group. (2011). Natural history collections as sources of long-term datasets. *TRENDS in Ecology and Evolution*, *26*(4), 153-154. http://dx.doi.org/10.1016/j.tree.2010.12.009

New York Botanical Garden. (2015, May 19). Google's Eric Schmidt & Sloan Foundation's Doron Weber on NYBG and World Flora Online. *Science Talk Blog*. Retrieved from http://blogs.nybg.org/science-talk/2015/05/googles-eric-schmidt-sloan-foundations-doron-weber-on-nybg-and-world-flora-online/

Park, I. W. (2012). Digital herbarium archives as a spatially extensive, taxonomically discriminate phonological record; a comparison to MODIS satellite imagery. *International Journal of Biometeorology*, *56*(6), 1179-1182. http://dx.doi.org/10.1007/s00484-012-0521-2

Schmidt, L. (2007). Digitization of herbarium specimens, a collaborative process. *Conference Proceedings of the Association of College & Research Libraries (ACRL) 13th Annual Conference*, 64-69. Retrieved from http://repository.uwyo.edu/libraries_facpub/7/

Shaffer, H. B., Fisher, R. N., and Davidson, C. (1998). The role of history collections in documenting species declines. *TRENDS in Ecology and Evolution*, *13*(1), 27-30. http://dx.doi.org/10.1016/S0169-5347(97)01177-4

Tegelberg, R., Mononen, T., & Saarenmaa H. (2014). High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. *TAXON*, *63*(6), 1307-1313. http://dx.doi.org/10.12705/636.13

Tulig, M., Tarnowsky, N., Bevans, M., Kirchgessner, A., & Thiers, M. (2012). Increasing the efficiency of digitization workflows for herbarium specimens. *ZooKeys*, *209*, 103-133. http://dx.doi.org/10.3897/zookeys.209.3125

Tingley, M. W., & Beissinger, S. R. (2009). Detecting range shifts from historical species occurrences: New perspectives on old data. *TRENDS in Ecology and Evolution*, *24*(11), 625-633. http://dx.doi.org/10.1016/j.tree.2009.05.009