Submitted on: 16.10.2017

Data In Libraries: The Big Picture
2016, August 10
Chicago, Illinois

Boundless Use of Indiana University's Biological Research Collections Possible in Partnership with IU Libraries

Jennifer Laherty
IU Libraries, Indiana University, Bloomington, Indiana, United States.
E-mail address: jlaherty@indiana.edu

Gary Motz
Center for Biological Research Collections, Indiana University, Bloomington, Indiana, United States.
E-mail address: garymotz@indiana.edu

ABSTRACT

In 2013, the Indiana University Libraries (IUL) embarked upon a rich partnership with the IU Center for Biological Research Collections (CBRC) to provide researchers and lay citizen scientists access to a curated, digital collection of 2D and 3D specimens, as well as critical scholarly data from researchers across the biological sciences at IU. We have implemented a Fedora 4 repository with Sufia 7 built on as a Hydra head to provide the CBRC with an extensible digital repository and well-curated digital archive. Success in this pilot program would facilitate support for the IU Libraries to undergo a major migration of existing digital collections – a migration that would boost digital collections onto the semantic web. Along the way, IUL and CBRC have realized each other's strengths and eagerly learn new protocols and methods for curating and preserving these institution-specific unique collections. Fundamentally, the desired outcome is to make the items in the CBRC abundantly available with the anticipation that their use will generate new discoveries while curating and preserving them for enhancement and future use. Furthermore, it is in this spirit of collaboration and a commitment to open technologies that we are aiming to extend these services more broadly to the entire IU community for digital scholarship and research. IUL regularly participates in community-based open source repository infrastructure engineering and global preservation networks. IUL's work with the CBRC is a natural extension of these commitments which benefit the library and informatics fields, both within IU and beyond.

Keywords: semantic web; biological research collections; research libraries; digital collections services; data repositories.

Introduction

In 2013, the Indiana University Libraries[1] (IUL) began looking for an extensible repository technology in order to update an institutional repository infrastructure that fails to provide depositors with a reasonably easy workflow experience, it doesn't extend easily to the multifaceted metadata schema needs of many different collection owners, and is not integrated with our digital collections preserved in a Fedora 3 repository. Discontinuing the development of the current institutional infrastructure has been an item on IUL's to-do list for several years. Additionally, the newly-formed Indiana University (IU) Center for Biological Research Collections[2] (CBRC) was gearing up for a major specimen digitization initiative that would provide object records, metadata, photographs, digitized specimen labels, and even 3D scans of close to 2 million collection objects. The CBRC is a research consortium that provides shared cyberinfrastructure and data management support for IU's natural history collections and is comprised of the faculty stewards of three major institutional repositories: the Indiana University Herbarium[3], the Indiana University Paleontology Collection[4], and the William R. Adams Zooarchaeology Laboratory[5]. These collections of physical specimens collected by researchers from IU's Departments of Biology, Geological Sciences, and Anthropology and their associated metadata (e.g. scanned PDFs of field notebooks, photographs, catalog data etc.) are being digitized *de novo*. Since the digital holdings of the CBRC are almost exclusively born digital content that has been digitized from physical holdings in very recent times, we believed that this collaborative effort would provide a unique framework for serving enhanced data services and digital preservation of institutional resources to a broader community at Indiana University.

IUL, a research intensive library system, maintains rigorous digital library collections practices in infrastructure, metadata, access, and preservation. The CBRC is a museum-type research center whose mission is to enhance collections-based research and education by providing its rich collections to global users, both digitally and physically preserved, in accordance to best library and archival practices in the areas of biodiversity science, botany, paleontology, zooarchaeology, and other areas. Since both the IUL and CBRC are natural and cultural heritage stewards, the symbiotic partnership takes a collegial and multidimensional approach to amplify traditional services managed as separate university units. At present, the CBRC is calling upon the IUL to create and deploy cutting edge technologies for archival, discoverability, and enhanced metadata synthesis that explores the value of augmented collection, both physical and digital, beyond present-day conventions. Furthermore, by collaborating with the CBRC, IUL seeks to understand the ways in which IU's unique and long-standing research collections may be delivered anew, breathing life into well-founded collections to drive exciting new scholarship and collaboration.

Desired Outcomes

Fundamentally, we are making the collection objects in the CBRC readily accessible, discoverable beyond the IU community, and well-curated with the anticipation that their use will generate new discoveries. Technological innovations and participating in the semantic web are two areas of mutual concern. Together, we have committed to several innovative technologies to enhance the preservation, discovery, and use of the CBRC digital specimens. Simultaneously, both the IUL and CBRC continue to engage the best current practices expressed in each domain-specific arena for their use on the semantic web. While no concrete plan has emerged for achieving this goal, we are in sync with the major contributors within the global biodiversity community.

Technological innovation

In this project there are four main repositories working together to promote the management, discoverability, and preservation of the digital counterparts of the physical objects cared for by the CBRC's member collections.

Sufia[6]. Sufia is a manifestation and extension of Project Hydra[7] which is a versatile technical framework built on the Fedora 4[8] repository. By using Sufia as a Hydra Head, this project gains a number of key functional repository features, such as: access controls (open access, closed/invited access, dark with no access unless catastrophic events occur, embargoed access at the discretion of rightsholders and collection owners); an API (Application Programming Interface) workflow that integrate with other workflows and services; and a more modern architecture, Rails, a model–view–controller (MVC) framework: Additionally, by extension of exploring Sufia for this project, IUL anticipates that this three-part technology stack will provide the basis for a new and better data service model for myriad collection owners at IU.

The IU Scholarly Data Archive (SDA)[9]. This is a tape-based distributed storage system and used to archive the original scans and photographs of the specimens. Access to the original high-resolution images is made available to researchers upon request, otherwise the web-based repositories offer a variety of derivative versions. The SDA is our robust archival solution for full-resolution original data as well as a backup of pertinent metadata from each of our digital asset management systems. Our Sufia repository is readily linked to the Scholarly Data Archive for deep access of digital archives.

Symbiota[10]. The Symbiota Software Project is dedicated to promoting collaborative infrastructure, automated workflows, and efficient digitization of biodiversity collections. This multi-institutional project site is a library of web-tools to aid biologists in digitization procedures and includes either open-source software implementations or best-practice recommendations for metadata acquisition, enhancement, or linking that includes: robust optical character recognition (OCR) paired with natural language parsing (NLP) and machine learning (ML), community based georeferencing tools (GeoLocate[11]), dynamic taxonomic checklists, and access control for the appropriate redaction and protection of sensitive data (e.g. rare and endangered species). Symbiota utilizes the Darwin Core Metadata Schema[12] as an interoperability language between individual collections and the biodiversity informatics community. Darwin Core is based on the standards developed by the Dublin Core Metadata Initiative[13] (DCMI) and is considered an extension of the Dublin Core specifically for biodiversity information.

Specify[14]. This is a National Science Foundation (NSF) supported, open-source software package (mySQL database system with Java frontend for operating system independent utility), which serves as a major discovery point for publishing specimen data from IU's natural history collections to the greater scientific community at large, loan-tracking tool, digital content management solution, and interface for browsing the holdings of each of the physical collections on the web or on an iPad. The Specify interface allows for us to interact directly with the digital object records in our Sufia repository to be called and directly linked to each specimen record.

Technology Discussion

IUL actively participates in the Fedora and Hydra developer communities to promote the use of open source repository infrastructure to meet the varied curation and preservation needs of cultural heritage organizations worldwide.

At its crux, the CBRC repository is built upon a Fedora 4 repository, which uses Resource Description Framework (RDF) triples to store linked data for all objects held therein, each with unique resource identifiers (URI). On top of Fedora 4 is Sufia, a user-facing application built as a Hydra Head.

The IUL is building this Fedora 4 repository to test various functional requirements to satisfy the CBRC. We automatically ingest daily scans, photographs, and metadata, and while this is not new and exciting, the IUL is presently building the ingest connections to receive these data direct from the Specify content management system. Regarding metadata, we were able to successfully configured the Fedora 4 repository to use both Darwin Core and Dublin Core metadata schemas in the item records. And we elected to use RDF rather than XML which is another advantage of Fedora 4 over Fedora 3. The storage of metadata in RDF is essential to our plan to participate in the semantic web (see discussion below). The CBRC specimens' rich metadata description was thoroughly scrutinized by IUL's metadata analyst librarian so that those descriptive elements are optimally designed for the semantic web integration and preservation.

The CBRC amplifies the IUL efforts by bringing new technologies to the infrastructure. The CBRC scans 2-D and 3-D objects and works in partnership with University Information Technologies Services (UITS) staff to expand upon open source visualizations tools which will be integrated with our Sufia repository for live manipulation of digital object at the point of discovery. The CBRC is also implementing augmented optical character recognition (aOCR) and natural language parsing (NLP) for the semi-automated transcription of handwritten labels on herbarium sheets and specimen labels. The CBRC is also undertaking georeferencing projects so that digitized maps of historic and scientific significance may be linked to the specimen data to provide greater locality based context for the collecting event metadata. Linking geospatial data, collection metadata, scans of handwritten field notes and observations, photographs, and research products is certainly the way of the future for the extension of pre-existing data into a multi-dimensional framework that ushers in the "big data" era in a very tangible way. For example, if a query of our Specify database pulled up a collection object record for a Devonian brachiopod (a ~400 million year old shelly fossil common in southern Indiana), we could see the full metadata for that specimen that includes collection date, collector name, collection locality, geologic age, etc., but we also have direct links to photographs of the specimen and its label, 3-D scans of the specimen, scans of field notebooks used by the collector when the specimen was initially collected, and much more content that dramatically enhances the contextual information and metadata for that particular specimen.

The CBRC Workflow

As biodiversity informatics and digitally augmented specimen-based research are new fields coming of age, the analytical techniques and research data that best afford a unique interpretation of the history, formation, and sustainability of biological diversity are held in collections. Many are the collections of private individuals and concerned citizens, some are the institutional holdings of universities and educational establishments, while others are entrusted to the public in museums by stewards of our natural history. As technological advances have improved our abilities to unravel genetic sequences, behavioral dynamics, ecosystem trophic structure, interactions among species and their environment, and numerous other ecological and evolutionary patterns and processes, some of the best methods of maintaining data accessibility and discoverability lie within partnerships forged in the indexed connectivity of the modern digital library. Here, we discuss the critical partnership between biodiversity informatics research scientists and the metadata stewards, research librarians, and digital collections specialists of the IUL.

This section provides an abbreviated technical overview of the digitization protocol developed by the CBRC for the digitization, archival, accessibility, and discovery of the treasure trove of biological data contained within the exceptionally well-curated collections of the IU Herbarium. These pressed plant specimens, mounted on acid-free paper with exceedingly detailed labels, record the history of botanical research in the state of Indiana, as the IU Herbarium is the official herbarium of the State of Indiana. Nearly 150,000 herbarium sheets enumerate the richness of data that physical plant specimens provide, along with critical metadata about the environment from which the plant was collected, plant species found in association with that collected specimen, geographic locality information, and a history of the taxonomic determinations that have been assigned to that specimen since it was first collected. In order to digitally preserve the contextual information for each specimen, associated research products (genetic sequences, field photographs, etc.) and to make the collection more broadly discoverable and accessible by researchers beyond IU, we embarked upon a comprehensive digitization scheme (also see Appendix 1 for a graphical representation) that will unfold in three parts: first, the construction of a novel information system to catalog pressed plant specimen metadata along with a digital photograph of each specimen; second, as specimen metadata is meticulously arrayed on a written label permanently affixed to each specimen, utilize optical character recognition (OCR) and handwritten text recognition (HTR) to record the metadata for each specimen in the skeletal information system generated by step one; and finally, to georeference each specimen and assign a geodetic datum (along with a margin of error radius) for each collection object. Many of these processes are fully automated, but numerous quality control checks (both in automated contexts and by humans) are required to ensure data integrity. Furthermore, with the tools provided by the Symbiota web portal, we can implement crowd-sourced and machine learning techniques to further enhance the proper parsing of metadata from OCR outputs into proper database entry fields. The linchpin that ties this entire workflow together is the digital repository built in Sufia 7. As numerous content types and bitstreams can be associated into one record, we now have a truly elegant solution for managing the comings and goings of continuously improving metadata.

This general workflow for building and enhancing a robust collection management and information system for the IU Herbarium is augmented and built upon by several other collections of natural history specimens, including the IU Paleontology Collection and the IU William R. Adams Zooarchaeology Laboratory. While we are presenting the two-dimensional object digitization scheme for clarity and to provide an appropriate scope of the overall project, the tremendous utility of the collaboration between the biological research collections and the library's digital repository system is in the robust integration of numerous digital collection objects. Since the CBRC is also actively digitizing entire skeletons using 3D laser scanners and fossil assemblages that are not appropriately categorized within a single 2D photograph, the broad linkages between numerous associated digital entities for each collection object is realized most fully when the Sufia object record bundles together each component allowing for full integration with open-source viewers of 3D content in addition to enriched metadata.

Linked Data and the Semantic Web

While this project allows the IUL to investigate the nimble nature of extending Fedora 4’s preservation capacity into any structurally sound research repository, of which there are many used by researchers/scholars at IU, the critical gain to be realized is exploring the extensible possibilities the RDF framework brings to the game. Fedora 3 only uses XML described objects, but by employing RDF with Fedora 4, we open our objects to a much wider world of use over the semantic web; every object will have its own URI and thus may be used on the web by anyone for any purpose. Taking advantage of Fedora 4 today allows us to utilize the latest RDF technologies currently in development so that we may readily serve web-ready content for development of interactive web services and broad dissemination and discoverability of dark data.

While RDF triples may be convoluted and tricky to read by many humans: "RDF is intended for situations in which information on the Web needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchanged between applications without loss of meaning."[15]

At present, the Fedora 4 CBRC repository includes URIs for the subjects and predicates in the RDF triple statement. We are cautiously assessing the best mechanism by which to include URIs for the object component of the triple statement and are participating in the conversation and development with the Fedora/Hydra community. Furthermore, we're engaged with other entities attempting to tackle these same issues of RDF implementation at scale. iDigBio[16] (Integrated Digitized Biocollections) is the United States' national resource for digitized information about vouchered natural history collections and is supported through funds from the NSF Advancing Digitization of Biodiversity Collections program. iDigBio's current *Guidelines for Managing Unique Resource Identifiers*[17] state that iDigBio utilizes HTTP GET requests to access web pages and metadata, using both RDF and XML. iDigBio is further evaluating a compact JSON-LD schema that preserves term (predicate) namespace, thus making them equivalent to URIs.[18]

Our CBRC team is particularly optimistic about the opportunity for data-enriched collections services to be innately more accessible and discoverable by means of the overall cyberinfrastructure in which they're served. The object-identifier environment and relational linkages that our Sufia repository provides better facilitates data aggregation by APIs that directly tie our RDF schema to parallel schemas for broad based interfaces for biodiversity informatics and other "big data" projects. National and international data aggregators, like iDigBio, the Global Biodiversity Information Facility[19] (GBIF), the WOrld Register of Marine Species[20] (WORMS), the Paleobiology Database[21] (PBDB), rely upon data curators and data stewards to continually provide links to digitized data; data that until the digital revolution has been lost to obscurity as "dark data" that is only useful, or even known, to the immediate managers of those data. These aggregators, and other more specialized resources like the Consortium of Midwest Herbaria[22], are the biological equivalents of WorldCat[23]: clearinghouses from which we can readily obtain data about specimen holdings and pertinent metadata for broad consumption.

Conclusion

IUL and CBRC demonstrate that together we amplify our power to consume and curate data in new and innovative ways. Behind the scenes of this quest are myriad library, museum, and archive staff, scholars, researchers, and lay citizens who take care to curate these assets. In addition, this is helping to breathe new purpose into the role of this subject librarian, who is not only learning the biodiversity landscape differently than through traditional collection development roles, but who is also promoting the use of the collection, albeit among librarians. This partnership receives the benefit of having two differently trained sets of staff co-curating a rich and unique institutional collection.

As we learn how these collections are used, we make better curatorial decisions and build better tools for using data. No longer is the discovery of such rich and powerful collections left to happenchance; rather, because of our due diligence regarding metadata, access, technology, and preservation, new generations of people can interact with these data and foster discovery and innovation. The symbiotic relationship between the human brain and the artificial intelligence of the semantic web are converging. The evolution and life cycle of our unique institutional collections are of utmost importance to curate using the best practices allowing linkages to flourish.

As the story of these biological research collections unfolds, we continue to learn more about how to extend university staff expertise to all IU collections. We recognize we are stronger by banding together our strengths. This collaborative effort reminds us that academic research libraries are incredibly well-poised to preserve digital scholarship, no matter which academic unit lays claim to the collection. It is gratifying to work with our CBRC colleagues who clearly value library expertise. Since IU contributes to important preservation initiatives (e.g. Research Data Alliance, The Digital Preservation Network, and HathiTrust), the CBRC trusts their library colleagues to give their collections excellent curatorial attention. Libraries generally work along-side the domain-specific subject arenas, in perhaps an odd agnostic role. Nevertheless, because of this, we have the potential to reach the largest audiences. The IUL and CBRC fully recognize that libraries are renewed in their mission to enhance scholarship, research, and education by serving academic institutions more broadly than ever before by taking advantage of partnerships with academic units.

ACKNOWLEDGMENTS

We acknowledge the leadership and foresight of those in the Indiana University Libraries and the IU Center for Biological Research Collections who made this project possible. Especially: David Polly, Robert McDonald, Jon Dunn, Kara Alexander, Julie Bobay, Will Cowan, Michelle Dalmau, Randall Floyd, Jim Halliday, Julie Hardesty, Nick Homenda, Eric Knox, Paul Rothrock, and Brian Wheeler.

REFERENCES

[1] Indiana University Libraries, https://libraries.indiana.edu/
[2] Indiana University Center for Biological Research Collections, http://www.iu.edu/~cbrc/
[3] Indiana University Herbarium, http://www.bio.indiana.edu/faculty/facilities/herbarium.shtm
[4] Indiana University Paleontology Collection, http://www.indiana.edu/~palcoll/
[5] William R. Adams Zooarchaeology Laboratory, http://www.indiana.edu/~zooarch/home.php
[6] Sufia, http://sufia.io/
[7] Project Hydra, https://projecthydra.org/
[8] Fedora 4, https://wiki.duraspace.org/display/FF/Fedora+Repository+Home
[9] Indiana University Scholarly Data Archive, https://kb.iu.edu/d/aiyi
[10] Symbiota, http://symbiota.org/
[11] Geolocate, http://www.museum.tulane.edu/geolocate/
[12] Darwin Core Metadata Schema, http://rs.tdwg.org/dwc/
[13] Dublin Core Metadata Initiative, http://dublincore.org/
[14] Specify, http://specifyx.specifysoftware.org/
[15] RDF Primer 1.1, https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/
[16] iDigBio, https://www.idigbio.org/

[17] iDigBio Guidelines for Managing Unique Resource Identifiers, https://www.idigbio.org/sites/default/files/videos/slides/iDigBio_URI_recommendation.pdf

[18] Alex Thompson, iDigBio Chief Cyberinfrastructure Architect. Personal Communication, June 2016.

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

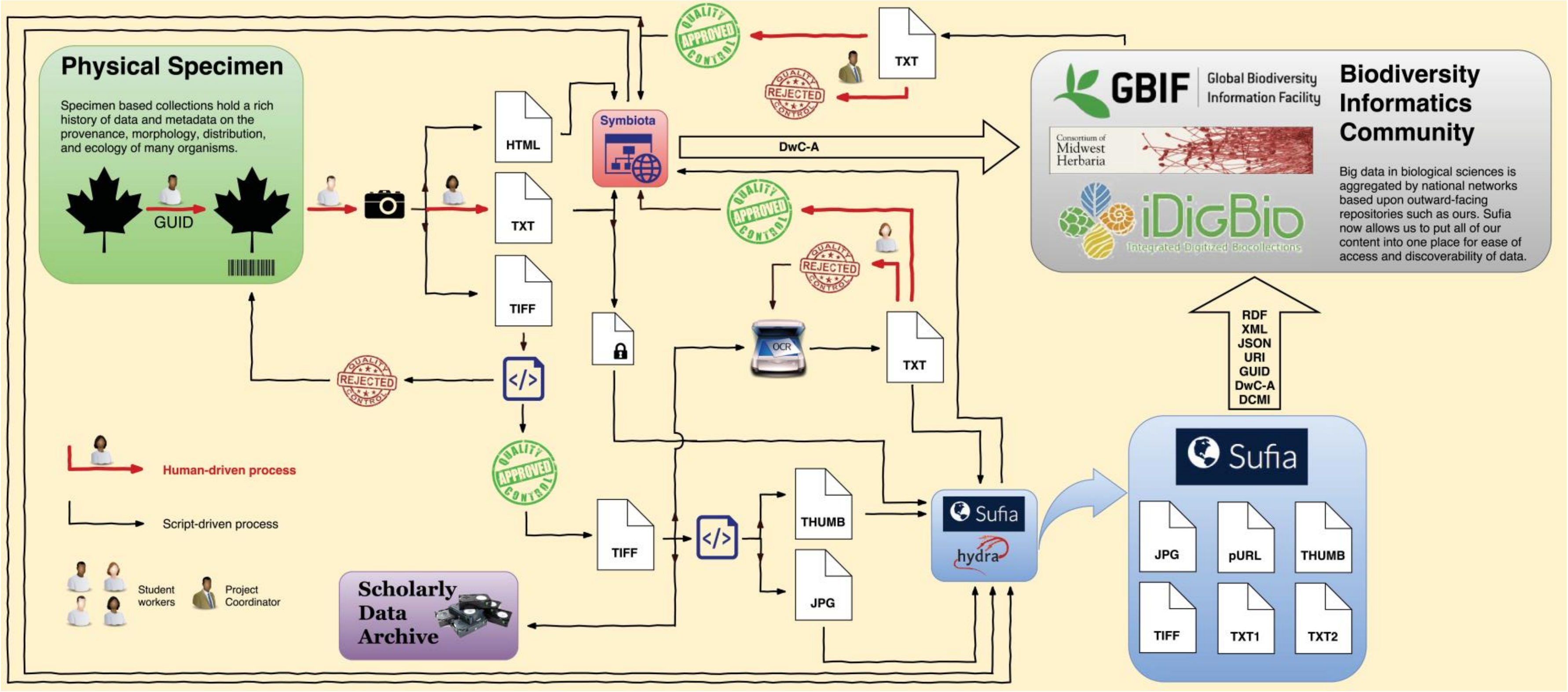
[20] WOrld Register of Marine Species, http://marinespecies.org/

[21] The Paleobiology Database, http://paleobiodb.org/

[22] Consortium of Midwest Herbaria, http://midwestherbaria.org/

[23] WorldCat, https://www.worldcat.org/

APPENDIX 1



Specimen based collections hold a rich history of data and metadata on the provenance, morphology, distribution, and ecology of many organisms.

Big data in biological sciences is aggregated by national networks based upon outward-facing repositories such as ours. Sufia now allows us to put all of our content into one place for ease of access and discoverability of data.