Data Curation Profiling of Biocollections

Bradley Wade Bishop
School of Information Sciences
The University of Tennessee
442 Communications Building
1345 Circle Park, Knoxville TN 37996
wade.bishop@utk.edu

Carolyn Hank
School of Information Sciences
The University of Tennessee
420B Communications Building
1345 Circle Park, Knoxville TN 37996
chank@utk.edu

ABSTRACT
In the contexts of the data deluge and open data, scientists studying biodiversity benefit from online access to global datasets of existing vouchered biological and paleontological collections. Using biocollections collected over time across the world allows for the advancement of scientific knowledge concerning evolution in process as well as species poleward migrations, an indicator of climate change. This study's purpose was to validate and expand the Data Curation Profiles (DCP) to digital biocollections and inform a DCP framework for worldwide biota. Ten biocollection producers, curating various types of specimens affiliated with the project building the United States' national biodiversity infrastructure, were interviewed using the DCP questionnaire. Results indicate there is extreme diversity in the curation of biocollections and additional DCP questions should be added to reflect the complicated approaches to biological data curation. Although discipline specific metadata creation tools, standards, and practices enable long-term sustainability of the U.S. digitization effort, some scientists would benefit from further clarification and guidance on the information needs of consumers beyond designated communities of expert users, and the long-term preservation of biocollections.

Keywords
Data curation, data curation profiles, Data provenance, Biocollections, Biology

ASIST 2016, October 14-18, 2016, Copenhagen, Denmark.

INTRODUCTION
For open science data to succeed globally, data curation relies on several essential roles: data authors, data managers and data users (NSF, 2005). The first are typically scientists supplying data, and the second are those who take on the role of data curation and management to ensure accessible, reliable, and reusable data. This is done on behalf of general or designated communities of consumers, the latter comprised the third role, "data users." No matter the data curation role, effective data curation planning and implementation benefits from a proactive, approach to understanding the "data story" in its active use environment; that is, at the onset and during research activities producing data, rather than as post-script.

Unlike some data, biocollections potential use and creation vary extensively. One factor complicating the curation of biocollections is the re-use potential. The potential consumers for many biological datasets goes well past the needs of scientists. The list of potential consumers for biocollections includes decision-makers and industry professionals related to agriculture, food security, public health, genomics, bioprospecting, ecotourism, mining, forestry, and several educational users from elementary through graduate education. The types and levels of data curation required to meet access and use expectations for these diverse and potentially extensive consumer communities requires extension consideration from the point of ingest, and the need for creating and migrating durable digital objects for deposit and storage to a digital repository, to value-added services for describing, understanding and creating tools to facilitate access, use, and re-use across biocollections potentially indefinite lifecycle of utility and usefulness. Further issues complicating access, use, and re-use is the potentially limitless, and yet unknown, groups of designated consumers that may benefit from the collections in not just the near but also the long-term future. Access controls (who is authorized to access and use the content) and intellectual

property concerns as to what might be done with the collections, particularly in reflection of a global information policy landscape, further complicates data curation policies, practices and procedures for biocollections.

Locality and genes are two other factors in the nature of biocollections data, which make their data curation distinctively complex. Georeferencing (e.g., assigning geographic coordinates) is the locality data of specimens at the point of collection. Georeferencing outside the built environment presents its own set of obstacles, but today global positioning systems (GPS) assist with ensuring accuracy of new data. Still, many legacy datasets, some new specimens, and fossils with locality shifting with continents over time, present challenges for curators of global biocollections ascribing accurate locality. The fuzziness of locality exists simultaneously in specimen records with the most naturally standardized metadata—genetics. Nucleotide sequences with specimen DNA assist with assigning species, but also variation within species.

Producer and consumer behaviors and preferences, and the significant and unique properties of the data itself create some specific problems for biocollections. This study's purpose is to validate and expand the Data Curation Profiles (DCP) Toolkit (2010), a methodology, including detailed data interview schedule, to allow data curators and manager to acquire an in-depth understanding the particular data curation needs of producers and their intended designated communities of end-users, also referred to herein as consumers. This study addresses a literature gap that will inform additional DCPs in the related biology disciplines and suggest refinements to the DCP framework when considering global multi-type biota. Although the study focuses on U.S. cyberinfrastructure, the approaches across biodiversity for locality and evolutionary taxonomy are universally applied making the study results applicable beyond one nation's approach.

Open data and geographic information policy

On May 9, 2013, the White House in the U.S. released the Executive Order, "Making Open and Machine Readable the New Default for Government." This Order stated that "Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth." This formal effort follows the White House's commitment to promulgate an "unprecedented level of openness in Government" in adopting three guiding principles: transparency, participation, and collaboration (Obama, 2009). In order for government data to be open, data must be collected, documented, organized, managed, and preserved. This follows a digital trend in US federal policy to make geospatial data coordinated and accessible across agencies through the National Spatial Data Infrastructure (NSDI) in order "to promote economic development, improve our stewardship of natural resources, and protect the environment" (Clinton, 1994, p. 17671). The cyberinfrastructure of the NSDI vision expands and echoes this in Executive Order 13642 (2013) stating that "openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth" (p. 244).

Prior to the Executive Order, the National Science Foundation (NSF) instituted required data management plans, and other agencies such as the National Institutes of Health (NIH) and the National Aeronautics and Space Administration (NASA) have followed suit. These plans have assured a higher degree of compliance with data sharing and preservation standards, but often, critical points in the data life cycle are still neglected. Data management planning benefits everyone by making data more discoverable, accessible, and usable. Open data exists in many arenas, but for geographic information and the particular geographic information of biocollections making data available is exceptionally complicated. Biocollections not only include specimens and/or images of those specimens, but locality data in various information types, complex data formats, ancillary files, dynamism, overall voluminous amounts, and most biota requires domain-specific metadata for appraisal and subsequent use. To organize the entire world of biocollections, standards were needed with attention to data provenance. Data provenance, also known as data lineage or pedigree, is critical to explain the history of data, linking it back to its original sources and thus completing the data life cycle. Issues related to provenance in digital biocollections relate to its sheer volume and immense scope and depth; failure to capture provenance at the time of the data's creation (legacy data) or during its manipulation in scientific models hamper scientific advancement. Failure to collect at the time of data creation makes it more difficult for scientists to determine the reliability of the data, validate results, and maximize the data's accessibility and reliability to duplicate studies (Higgins, 2015). Producers as an essential first line of defense in digital preservation is a long-standing, fundamental principle (Garrett & Waters, 1996). Hence, the imperative for effective data management, and the requisite planning and understanding to facilitate this, is essential for managing biocollections in their primary, secondary and preservation use environments.

Biocollections Background

Biodiversity presents one example of why open data with adequate data provenance is required to advance science. Modern biodiversity research stems from Joseph Grinnell, a Darwinian naturalist that studied how environmental changes drive the process of organic evolution at the University of California Berkeley (Griesemer & Gerson, 1993). In Molecular Biology, scientists usually study organisms with a short development period to track changes in controlled generations (Bouldin, Snelson, Farr, & Kimelman, 2014). Grinnell mandated a rigorous and standardized methodology of specie collection to represent the natural world in great detail to study evolution in action (Ilerbaig, 2010). The Museum of Vertebrate Zoology, established in 1908, allowed scientists to collect,

label, and preserve specimens with field notes, field maps, and sometimes photographs, and through systematized processing creating an institutional collection that enables study beyond one individual or one team of scientists' lifetimes.

Biocollections have historically been housed in university collections (e.g., natural history museums), federal collections, discipline-specific collections (e.g., herbaria), private collections, and zoos. Although many follow the Grinnell method, this is not the case universally. By digitizing biocollections around the world, the data and images of millions of biological specimens can now be made available in electronic format for the research community, government agencies, students, educators, and the general public. Sharing biocollections digitally not only advances the study of evolution, but also captures poleward range shifts of species at a scale that provides another indicator of climate change (Hampe, Arndt, & Petit, 2005). Scientists are predicting that future species declines will approach historical mass extinction levels within this century. For these various scientific endeavors data must be collected, shared, and preserved in a systematic way as each specimen collected present one snapshot in time of archiving the natural world.

The National Science Foundation's (NSF) Advancing Digitization of Biodiversity Collections (ADBC) addressed these science needs through the iDigBio project (https://www.idigbio.org/). The project's mission to develop a national infrastructure that oversees implementation of standards and best practices for digitization; builds and deploys a customized cloud computing environment for collections; recruits and trains personnel, including underserved groups; engages the research community, collections community, citizen scientists, and the general public through outreach activities; and plans for long-term sustainability of the national digitization effort. The project does provide data ingestion guidance and the iDigBio portal, as of July 24, 2016 contains 64,015,275 specimen records and 14,320,405 media records. Clearly, the prescriptive approach has led to successful contributions from many biocollections. Perhaps, using DCPs could assist biocollection managers to capture information that informs data curation with a focus on the various stakeholders beyond the technical requirements for data ingestion. In several cases, the biocollection managers are one step removed from the scientists, citizen scientists, and so on that collected the specimens.

Data Curation Profiles

A data curation profile "captures requirements for specific data generated by researchers articulated by the researchers themselves" (http://datacurationprofiles.org/purpose). Through interviewing the data producers themselves (a.k.a., the "first line of defense," it allows essential data provenance information to be captured, such as why and the how the data is created, how the data evolves over various stages of the research enterprise (i.e., raw, processed, analyzed, and so forth), and the resulting data types. Through enhanced and detailed understanding of the "data story" directly from data producers, data curators can better prepare the data, manage it, and design value-added services, such as enhanced metadata application and analytical tools, for future users and long-term preservation.

With the attention paid to "big data," and now even "small data," in recent years, data science has emerged as a critical area of professional and research practice. Data curators can be seen as an evolved role where the traditions of librarianship and information sciences focus on the intersection of information, people and technology, have merged with the more nascent practices of data analytics and visualization. In the last decade, many academic librarians have explored options in providing data curation services and resources to producers and consumers. Datasets across scientific fields vary widely and diverge in information representation practices from materials traditionally collected by academic libraries (Carlson, 2010). Although a strong argument has been made for academic libraries to house data on campuses, this study uses the tool developed by academic librarians to build DCPs for biocollections regardless of the organizational entity and individuals ultimately charged with data curation as future cyberinfrastructure may not be directly tied to any traditional information agency per se.

The challenges of biocollections extend not from the born digital collections, with typical challenges associated with technological obsolescence (e.g., format types, software applications etc.), it also extends to reborn or yet-born digital data. Digitizing the variety of biocollections has been difficult given data vary from birds in boxes to pressed vegetation to insect fossils. The "data deluge" described by Hey and Trefethen (2003) has become accepted as a fundamental characteristic of science today as scientific data continues to increase at a rate of around 30 percent per year (Pryor, 2012). Living organisms propagate more quickly than any potential data collection and the data has been big since the dawn of life itself. In today's data intensive science climate, it is becoming more widely realized that in order for data to be discoverable, accessible, and usable, both now and over time, it must be collected, documented, organized, managed, and curated (Strasser, Cook, Michener, & Budden, 2012). Data provenance must be captured at the time of data collection and ingest using Grinnellian method or some other highly detailed method, to prevent information entropy from limiting future use of the biota in the aggregate (Michener, Brunt, Helly, Kirchner, & Stafford, 1997). Informed data curation beyond the scientists' self-archiving and personal information management is necessary to meet current and future consumer needs of biodiversity data. Although many data lifecycle models exist, the terminology for the Data Curation Centre's Curation Lifecycle Model is most useful for this study (Higgins, 2008). One ethnographic study of researchers indicated a data life cycle truncated when scientists did not consider steps related to

data dissemination, data deposit, data preservation, data discovery, and data repurposing as part of their data life cycle (Jahnke & Asher, 2012). Perhaps, the DCP framework developed from the information agency perspective can provide insights needed by biocollection data producers, managers and curators to fulfill the breadth of user needs related to the preservation, storage, access, use, and reuse, and transformation elements of the lifecycle. In addition, future DCP questions for biological data curation could reflect some of the specific tools, standards, and approaches common in the field of biodiversity.

METHODS

This study adopted a qualitative, semi-structure interview approach, derived from the DCP methodology. The sampling frame comprised 33 of the Thematic Collection Networks (TCNs) collaborators affiliated with iDigBio (e.g., Fossil Insects). An effort to reach all the 15 different types of TCNs was made to validate and expand DCP framework across biodiversity, but a member from each did not participate. Each TCN has between 7-to-92 collaborators depending on the number of entities collecting the different specie types. Through attendance at an international conference of biocollection managers, a snowball sampling approach was used to reach the 'data people' behind all these projects. Each project has a variety of approaches to which team members served in the role of data manager (e.g., the scientist themselves, an IT person, graduate assistant). Therefore, each potential study participant was recruited through this purposive sampling technique based on if they were tasked with data curation for their project. The scale of projects varied from entire natural history collections to specific researcher datasets.

From snowball recruitment at and post-conference, ultimately ten phone interviews with biocollection managers were conducted over a six week period based on participant availability. In the DCP methodology, the interview schedule provided captures step-by-step data provenance information that scientists require to determine fitness for use in future projects. This includes perceptions on the degree to which a dataset is suitable for a particular application or purpose, encompassing factors such as data quality, scale, interoperability, cost, data format, and so on. Again, locality and evolutionary taxonomy are two problematic metadata elements specific to biocollections. The interview schedule was derived from the DCP methodology to account for particularities of biocollections. This was intended to enhance and better target the population under investigation, building the established DCP methodology, which itself was informed by literature, interviews with scientists, and validation from a panel of expert reviewers (Witt, Carlson, Brandt, & Cragin, 2009).

The interview schedule for this study, as with DCP schedule, contains questions related to how the data is collected and other critical data attributes, including size, format types, organization, description and representation, and storage. Deviating from the DCP methodology, since data ingest was not an actual, or applied, outcome, interviewees were not prompted to answer questions on ingest and deposit as it was not relevant for the objectives of this study. The interview schedule consisted of the following questions:

1. Please provide a brief overview of the research associated with the data we will be discussing.
2. Description of the data
    a. Approximately, how many data files exist?
    b. What is the average size of the data files? (units) and/or overall (total file size)
    c. What format(s) are the data stored in?
3. Data Flow & Use
    a. How was the data acquired/collected?
    b. What specific software programs or tools/hardware were used in the collection/generation of the data?
    c. How was locality determined? Place/time?
    d. What specific software programs or tools/hardware are required to utilize this data?
    i. Describe briefly the way the data is currently organized (i.e., file name conventions, existing metadata, units, and so forth)?
4. Storage
    a. Where are the files currently stored? Include the storage media(s) and any tools used in your management of the data.
    b. Are there backups of the data?
    c. Who is primarily responsible for managing these files?
5. Stakeholders
    a. Who is the intended audience of this data? Is the data intended to be made available for re-use by others?
    b. Who might you imagine would be interested in this data? (e.g., other researchers in my field, researchers outside of my field, practicing professionals, policy makers, etc.)
    c. How might this data be used by these people?
6. Jobs
    a. What is your current job title?
    b. How many years in total have you been working in your current job?
    c. How many years in total have you been working with biocollections?
    d. Describe your work setting?
    e. Please indicate your credentials and degrees.
    f. Please provide any other educational or training you have received that is applicable to performing your job.

The ten completed interview were recorded and transcribed. The transcriptions were analyzed using NVivo. Grounded theory application of open, axial, and selective coding was

applied to generate categories and broad themes across responses to the questions.

RESULTS

Results from the interviews were threefold: (1) DCPs of biocollections are diverse; (2) the DCP questionnaire would benefit from greater specificity when used with biocollection managers; and (3) managers of digital biocollections have different conceptualizations of stakeholders, including data authors, managers and users, and their information needs, and the long-term preservation of biocollections. Further, the wide variety of job titles, work settings, education and training present some considerations for curriculum development and job analyses in these areas. The following results are subsectioned as ordered in relation to the derived DCP interview schedule used in the study.

Diversity in biodiversity research

Overall, research in biodiversity concerns the number of different organisms in an ecosystem, but also variation within a species and between species. To study research questions in these areas globally and throughout time, scientists need access to both neontological and paleontological specimens in various biocollections and DNA banks and genetic resource repositories. One purpose of all biocollections is to prepare and present data to allow the biodiversity community to store, access, use, and exchange data. Managers characterized the overview of the biodiversity research being done with their biocollections at differing scales depending on their proximity to the biological materials that become data.

The data manager responsible for aggregating data—henceforth referred to as aggregators—across biocollections have a workflow that is removed from the physical specimens and give a research overview that was broad. The three aggregators interviewed manage Animalia, Plantae, Fungi, Protista, Protozoa, Bacteria, and Protoctista, and all that data facilitates the study of "evolutionary biology" or "climate change" or challenges in "retrospective georeferencing". Throughout the DCP questionnaire, aggregators discussed data science issues that constitute a research overview, but since that knowledge work enables other science to occur these were not mentioned in responses to the research associated with the data itself.

Still, most data managers that are trained scientists gave responses more specifically related to biodiversity research conducted using their types of biocollections. The three herbarium managers mentioned how researchers could study the variety in plants across the southeastern U.S. using their collections, or the angiosperms across a state over time, or distribution of lichens and bryophytes across the entire U.S. The two data managers of fish collections knew that their collections had been used to study past and present trends in biodiversity and genomic traits of fresh and marine species with spatial and temporal context across the southeastern U.S. and the Pacific and Indian Oceans.

The two paleontologists had fossilized species on different ends of the Animalia spectrum. The paleoanthropologist with data for hominids in the Plio-Pleistocene geology of east Africa facilitates the study of human origins. In contrast, the invertebrate paleontologist biocollections allow for a better understanding of changes overtime to trilobites and other species.

Diversity in data

The overview of research from the data managers indicates the global and near limitless bounds of biodiversity and the bottles of deep-sea fish, pressed plants, and hominid remains still under excavation in the rock, all feed into larger datasets that enable biodiversity research. The number of files and average size were difficult questions for many data managers, but some could provide exact counts. As this study did not include any planned data ingest, when participants did not know they were not pressed. For example, many data managers did not know or give data extent/size estimates. To effectively answer, two data managers would need to open up the database and gave "don't know off the top of my head" responses. Two data managers provided approximations of "about ten thousand images" or "about a half a million unique geographic points". Three other data managers were definitive (e.g., "six hundred and eighty thousand specimen fish collection," "twelve hundred specimens in it" and "50 million specimens in the southeast"), but admit that collection size is a "bit of a moving target" with new discoveries continuously adding to the biocollections. This moving target aspect is most evident in the primary aggregator for iDigBio who indicated there were 44 million specimen records ingested into the database. The portal also held 44.6 billion text-based records and 12.3 million media objects (e.g., pictures, sounds, 3D scans). The two other aggregators would not be able to answer the question as one uses Amazon Relational Database Service (RDS) on big data stacks to extract content on very large aggregates of data, and the other aggregators also work across large scale datasets without exact figures.

For data managers that did not address the number of files, average file size was an equally meaningless question. Three did not know. Five guessed average sizes of files like TIFFs and PDFs in the dataset, or provided a total collection guess "it's all under a terabyte.," or a piecemeal estimate "about a megabyte for every ten records". The data formats did not present the same challenges as the formats inform use and follow community standards. Images were indicated to be mostly JPGs, but high-resolution TIFFs, CR2s, and NEFs were also mentioned as well as 3D scans. Image servers connect that data with other records such as locality and taxonomy field notes for specimens available in PDFs. Other data formats mentioned included Excel, SQL, PostgreSQL, and PostGIS. Darwin Core was mentioned by five of the data managers as the metadata format the data was stored in.

Diversity in flow, use, and storage

Data flow and use differs at different levels of aggregation and data managers answered these questions from their idiosyncratic context of the data lifecycle. The aggregators ingest smaller biocollections across TCNs and other scientists seeking to publish data to make it available for others to use. Only one aggregator elaborated on how the data was originally collected explaining the written descriptions in field notes that followed the Grinnellian method and how those are building blocks that support future science to occur. Even before data ingest, data collection decisions and practices impact provenance, data quality, and limits data's future use. Other data managers acknowledge a dichotomy in their collections—data collected and data acquired. The other data managers point out that data collected with pen and paper, or nowadays on an iPad, follow a variety of field note methods to live in biocollections alongside other data acquired. Data acquired in biocollections comes from data prior to scientific data collection standards, gifts of private collections, and amateurs' donations. In some instances, those items go into teaching collections and not for research use.

The aggregators listed tools and hardware to use in the collection and generation of the data. The primary aggregator is using custom solutions and homegrown tools built for only the main portal. Another aggregator mentioned the Field Information Management System (FIMS) that allows a specimen gathered in one place to remain linked to all the data resulting from it (e.g., the specimen, parasites that were collected off of it, genetics samples, and so forth). Specify was another tool mentioned by an aggregators (http://specifyx.specifysoftware.org/), but across authors was mentioned by five of the ten interviewed. Specify is a database platform made up of 143 tables and 2,400 fields of information that takes data and parses fields to track specimen records, assign URLs to images on web servers or saves a copy of the image online that links to specimen records, and publishes data online. Two data managers mentioned Symbiota (symbiota.org). Symbiota supports entire specimen digitization and collections management workflows plus extensive data exchanges with other systems for virtual floras and faunas (Gries, Gilbert, & Franz, 2014). One data manager mentioned developing a similarly functioning workflow for Android.

The location of these specimens is a key factor in the study of biodiversity. The way locality has been determined impacts future use. To review, a locality is a point of reference (e.g., coordinates) on a particular geographic coordinate system (e.g., Universal Transverse Mercator (UTM)). Some managers assign a locality a unique identifier within their collections (e.g., Hillier Canyon five). Place names associated with exact locations assist users and mangers that do not think in coordinates. Due to the variety of ways locality is measured, aggregators are "agnostic about locality data" as ingested data has to have some quality assumed despite known complications. All locality data should be georeferenced and conform to Darwin Core standards for iDigBio ingestion. Newer data benefits from the use of a GPS, but legacy data does not have that benefit. Legacy data locality often includes a string of text that links a specimen location back to a known geographic reference point in the built environment (e.g., Missouri River 3 mi. SE of Pierre, South Dakota, USA, South Dakota, Hughes). To assist with determining locality for legacy data, GeoLOCATE (http://www.museum.tulane.edu/geolocate/) was developed to calculate a string of text into a point (i.e., X,Y) and a representation of uncertainty (i.e., polygon of where the specimen might be within a radius of the point limited by terrain). GeoLOCATE was mentioned in 7 out of 10 interviews. For paleontologists, a combination of stratigraphy and topography assists with estimating a specimen's point in time as well as space. Programs through FishNet2 (http://www.fishnet2.net) exist to help improve the locality of legacy fish collections performing similar functions.

Data managers mentioned typical software programs to utilize the data, including Microsoft Access and Excel, R, ESRI ArcGIS, QGIS, DIVA-GSI, and PostGIS, but as one aggregator rightly indicated: "That's a giant question, I mean you have got to be kidding?" However, for the DCP questionnaire it is beneficial to list some of the typical software to allow for future user, at the risk of potentially confounding interviewees with either too limited or, more typically, too extensive of an application inventory from which to choose.

The information organization of the biocollections revealed some commonalities in practice. Data managers use data dictionaries to ascribe scientific names to species (e.g., evolutionary taxonomy dictionaries), locality, sequential unique identifiers for barcodes and databases, and time stamps to log any changes to the data and by whom. Unlike bibliographic classifications, each curator uses their own collection-specific method for information organization that reflects convenience of use for data managers and users in naming conventions. The standardization across biocollections is closer for metadata with five out of 10 mentioning Darwin Core.

The data is nearly all stored on servers in the cloud. Backups occur through associated state, university, organizational, or commercial cloud service servers. The majority of the individuals responsible are from computing or IT at the data managers' respective institutions. Only the aggregators and one smaller collection rely on the data manager to do the backup.

Diversity in stakeholders

The stakeholders presented difficulty for some managers to conceptualize. There is a dichotomy between those that say "our answer is always everyone" to more specialized list of biodiversity researchers, reflective of the designated community of users in reference to the OAIS Reference

Model (2012). The data authors may be ecologists, taxonomists, systematists, and anyone working in biology in the natural sciences, as may be their respective designated communities or data users. While the data authors and users are the most typical and heaviest users of the biocollections in order to advance knowledge across their diverse disciplines, there are other key stakeholder groups to consider. For example, this evolving knowledge informs funding decisions and policy developments crafted by decision-makers from non-governmental organizations, state and federal agencies, and others making choices related to environmental impact assessments, public health risks, or other plans related to habitat protection, restoration, and mitigation. As such, these stakeholders play a vital role in the data curation needs of biocollections.

And while efforts can be made to select, appraise, and ultimately manage data that are deemed utile for current stakeholder communities, data authors, and data managers also need to be aware of not-yet-known communities of future users, and acknowledge the diverse and potentially extensive uses these collections may be put to use. Stakeholders also extend well beyond these essential policy-makers and funding agencies. Ultimately, elementary to graduate school educators, their students, and any visitor to a natural history museum could benefit from access to these biocollections by learning about the biota around them. Other potential data user stakeholder groups include amateur scientists, citizen scientists, and other enthusiasts. As a result, it appears to be a futile exercise to identify all stakeholders, as an appropriate response could be everyone. Rather, the challenge lies in how to prioritize select stakeholder groups over others, and their specific needs, when considering not just the technical and organizational investment in effective life cycle management of digital biocollections, but also value-added tools and services to enhance access, use and re-use both now and into the future.

Diversity in jobs

The final six interview questions for the data managers asked demographic questions related to their job title, years of professional experience, work setting, education, and training. The job titles varied, including IT Expert, Workflow Coordinator, Post-Doc, Collections Manger of Ichthyology (2), Invertebrate Paleontology Collection Manager, Curator of the Herbarium, Curator of Biodiversity Informatics, Associate Professor in Anthropology, and Information Architect. The time in their respective current job ranged from 20 years to only a few months, but the time these managers had worked in biocollections averaged 20 years with only two with less than 10 years' experience. The work settings included typical office spaces, but every individual referenced their reliance on computers and servers to process and manage data. There are specimens, cameras to document the specimens, and lab equipment for preparing specimens that undergird all these digital efforts, but gone are the days where managers "were considered a glorified bottle shufflers […] now there's been this digital revolution, where now we spend probably eighty percent of our time sitting in front of a computer working with data, augmenting data, entering data, reporting data to the outside world."

Eight of the ten mangers had education in biology, with four having a Ph.D. in biology, one with a Ph.D. in anthropology, and six holding a master's in biology as well. Due to the information technology skills required to do the data curation of digital biocollections, two aggregators did not have an educational background in biology. One had an associate's degree in political science and the other a bachelor's degree in physics and medieval literature. Clearly, the knowledge, skills, and abilities to perform these tasks were not part of the biology curriculum and required additional training for all data managers. Nearly all indicated that they taught themselves analytic tools like R and GIS and programming like SQL, Java, and Python. Three of the ten managers mentioned the value of learning from other managers through professional societies or visiting others' collections. Two managers indicated helpful workshops on workflows and georeferencing from iDigBio, and one mentioned the value of Taxonomic Databases Working Group (TDWG) workshops. In this emergent area, reliance on informal community training takes precedent over more formal, degree based education and curriculum.

CONCLUSION

The study informs the data curation profile framework by using a derived DCP questionnaire with digital biocollections. This study also validated the DCP, but provides some areas to expand and be more specific when creating profiles for biocollections. As outlined in the results section: (1) DCPs of biocollections are diverse; (2) the DCP questionnaire would benefit from greater specificity when used with biocollection managers; and (3) managers of digital biocollections have different conceptualizations of stakeholders, the information needs of stakeholders, and the long-term preservation of biocollections. One limitation of the study was snowball sampling that led to only certain types of biocollection managers being interviewed. Aggregators of multiple biocollections face different data management issues as any data aggregated must meet stringent guidelines before being ingested into a digital repository for access, use, and re-use. For the DCP framework, the interviews with ichthyologists and botanists were most useful as they could inform the data needs, uses, and storage of those biocollections. Future research should include interview subjects representing Animalia (e.g., Mammalia), and Plantae, as well as others (e.g., Fungi, Protista, and Protozoa).

Recommendations for future use of Data Curation Profile questionnaire in biocollections

Although the study focuses on U.S. cyberinfrastructure, the approaches across biodiversity for locality and evolutionary taxonomy are universally applied making the study results applicable beyond one nation's approach. The saturation of common tools for represented biota across U.S. collections

indicated that data curators should familiarize themselves with these tools and add those data creation and use tools to DCPs. In addition, to use the DCP questionnaire the number of files and file sizes needed to be more specific. Any image, PDF of field notes, locality form, loan form, and other ancillary files for a specimen could be considered individual files. Furthermore, a file may be at a collection level (e.g., lot) and not an item level (e.g., fish). A lot is where "all of the specimens of the same species collected at the same time by the same person are all amalgamated into a single jar and given a single catalog number." The file number and size could not be addressed by most collection managers because of the volume and dynamism of biocollections.

Aggregators of data should receive unique questions as data curators of museum collections and similar sized projects face different issues. Aggregators already have ingested data, but perhaps the DCP could help them think through the other parts of their data's lifecycle in greater detail as relate to data dissemination, data deposit, data preservation, data discovery, and data repurposing. If managers intend to share data, then none of the DCPs questions should go unanswered. The results indicate that some areas of the data lifecycle require further reflection not only by individual managers, but by the field of biodiversity research as a whole.

Data Curation Education

The education and training questions for biocollection managers indicate that aggregators of biocollections did not gain data curation education formerly, but relied informal training, building their own tools and benefitted from a computer science and information technology background and experience. As one noted: "All of this was just sort of like, Wild West. No road maps". Although this cutting edge aggregator was referring to his training in early 1990s web development, the same is reflected in the current state of data curation in biocollections. Perhaps, informatics should be taught throughout biology and related disciplines. The lengthy incubation period required in the extensive education held by many managers in this study (i.e., graduate work, post-doc) point to the value of domain knowledge. Whether formal data curation education would benefit these individuals is speculative, but by using the DCP clear gaps emerge in aspects of the data lifecycle not fully grasped that if addressed would benefit all biodiversity research.

REFERENCES

Bouldin, C. M., Snelson, C. D., Farr, G. H., & Kimelman, D. (2014). Restricted expression of cdc25a in the tailbud is essential for formation of the zebrafish posterior body. *Genes Dev, 28*(4), 384-395. doi: 10.1101/gad.233577.113

Carlson, J. (2010). The data curation profiles tookit: Interviewer's manual. *Data Curation Profiles Toolkit.* Paper 2. http://dx.doi.org/10.5703/128828431565

Clinton, W. (1994, April 13). *Coordinating geographic data acquisition and access*: *The National Spatial Data Infrastructure*. Executive Order 12906. Retrieved from http://govinfo.library.unt.edu/npr/library/direct/orders/20fa.html.

Consultative Committee for Space Data Systems (2012). *Reference model for an Open Archival Information System (OAIS).* [Recommended practice, CCSDS 650.0-M-2, Magenta Book). Washington, D.C: CCSDS. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Exec. Order No. 13642, 3 C.F.R. (2013).

Garrett, J., & Waters, D. (1996). Preserving digital information: Report of the Task Force on Archiving of Digital Information. Washington, DC: The Commission on Preservation and Access and RLG.

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

Griesemer, J. R., & Gerson, E. M. (1993). Collaboration in the Museum of Vertebrate Zoology. *Journal of the History of Biology*, *26*(2), 185–203. Retrieved from http://www.jstor.org/stable/4331259

Hampe, Arndt, & Petit, Rémy J. (2005). Conserving biodiversity under climate change: the rear edge matters. *Ecology Letters*, *8*(5), 461-467. doi: 10.1111/j.1461-0248.2005.00739.x

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation, 3*(1), 134-140. Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48

Higgins, S. (2012). The lifecycle of data management. In G. Pryor (Ed.), *Managing Research Data* (pp. 17-45). London, UK: Facet Publishing.

Ilerbaig, J. (2010). Specimens as records: scientific practice and recordkeeping in natural history research. *The American Archivist*, *73*(2), 463-482.

Jahnke, L., & Asher, A. (2012). The problem of data. Washington, DC: Council on Library and Information Resources (CLIR).

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). Nongeospatial metadata for the ecological sciences. *Ecological Applications*, *7*(1), 330-342.

National Science Foundation, National Science Board. (2005). Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. http://www.nsf.gov/pubs/2005/nsb0540

Obama, B.H. (2009, January 21). Transparency and open government. Memorandum for the Heads of Executive Departments and Agencies. Retrieved from: http://www.whitehouse.gov/the-press-office/transparency-and-open-government

Pryor, G. (2012). Why manage research data? In G. Pryor (Ed.), *Managing research data* (pp. 1-16). London, UK: Facet Publishing.

Strasser, C., Cook, R.. Michener, W., & Budden, A. (2012). Primer on data management: What you always wanted to know. UC Office of the President: California Digital Library. Retrieved from: http://escholarship.org/uc/item/7tf5q7n3

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing Data Curation Profiles. *International Journal of Digital Curation* *4*(3), 93-103. doi:10.2218/ijdc.v4i3.117