Grinnell to GUIDs: Connecting Natural Science Archives and Specimens

Christina Fidler, Museum Archivist, Museum of Vertebrate Zoology, University of California, Berkeley

Barbara Mathé, Museum Archivist and Head of Library Special Collections, American Museum of Natural History

Rusty Russell, Program Director for Collections & Informatics, United States National Herbarium, Smithsonian Institution

Russell D. "Tim" White, Director of Collections and Operations, Peabody Museum of Natural History, Yale University

Abstract

Four recipients of Council on Library and Information Resources (CLIR) Hidden Collections grants in natural science museums—two archivists and two scientists—discuss their view of and approach to access for resource discovery, the effects of digitization, integrated access in aggregators, and the issues of long-term digital preservation. The long-term goal is to integrate and link access to digital data and images across natural science institutions, whether libraries, archives, or museums.

> "It is quite probable that the facts of distribution, life history, and economic status may finally prove to be of more far-reaching value, than whatever information is obtainable exclusively from the specimens themselves."
>
> —Joseph Grinnell (1910, 166)

Joseph Grinnell, the first director of the Museum of Vertebrate Zoology at the University of California, Berkeley, foreshadowed—by more than a century—the growing recognition within today's natural science community of the value of all information found and gathered during a collecting event.

In 1908, Grinnell developed and implemented a detailed protocol for recording field observations. These integral materials, gathered along with specimens, contain extensive information that may not appear on labels attached to or stored with collections objects. They may include detailed accounts of individual species' behaviors; annotated topographic maps; photographs of collecting sites; observations that did not result in specimen collection; interactions with local or indigenous populations; and other data, such as weather conditions, vegetation types, vocalizations, and various evidence of animal presence in a given locale.

Integrated Digitized Biocollections (iDigBio), part of the National Resource for Advancing Digitization of Biodiversity Collections (ADBC) funded by the National Science Foundation, is an aggregator that will allow data and images for millions of biological specimens to be made available electronically. In March 2014, iDigBio, Yale University, and the Field Book Project sponsored a workshop focused on digitizing original source materials associated with scientific collections. This gathering reflected a growing momentum toward providing access to all types of resources related to natural science, not only specimens and species publications but field recordings—in the form of books, notes, sketches, correspondence, and audiovisual materials along with records of the research conducted using these things. Although there is still a crying need for the time-consuming work of processing archival collections, the long-term goal is integrated access to natural science information—data and images—wherever they are held. As the development of linked data applications using globally unique identifiers (GUIDs) proceeds, this goal can become a reality.

▶ Christina Fidler, Museum Archivist, Museum of Vertebrate Zoology, University of California, Berkeley

Project title: Cataloging Hidden Archives of the Museum of Vertebrate Zoology: Increasing Integration and Accessibility for Interdisciplinary Research (2011)

Access (resource discovery): Why and how have you made it possible for your archival collections to be found by those looking for them (or perhaps by those who did not know they existed)?

Beginning in 2003, the Museum of Vertebrate Zoology (MVZ) at the University of California, Berkeley, executed the Grinnell Resurvey Project, a large-scale resurvey of ecological transects in California. These transects define different specified areas where the abundance and distribution of faunal populations can be measured. To conduct this extensive resurvey, researchers relied primarily on the archival field notes that document the floral and faunal conditions of the same transects roughly 100 years ago. It was an astonishing testament to the importance of these historical data. The resurvey was a catalyst for the museum's CLIR Hidden Collections grant, which began in 2012. I am the archivist hired on the CLIR grant, which is dedicated to establishing a formal repository and cataloging the archival field note collections. I work under the supervision of the grant's co-principal investigators, Carla Cicero and Michelle Koo.

In 1910, Joseph Grinnell, the museum's first director, wrote with alarming accuracy, "At this point I wish to emphasize what I believe will ultimately prove to be the greatest value of our museum. This value will not, however, be realized until the lapse of many years, possibly a century, assuming that our material is safely preserved. And this is that the student of the future will have access to the original record of faunal conditions in California and the west, wherever we now work" (Grinnell 1910, 166).

As mentioned earlier, Grinnell developed a means for capturing data while collecting and observing in the field. The Grinnell method of taking field notes has been practiced over the last century and is still being practiced much as it was originally at the MVZ. This has resulted in more than 700 volumes of field research materials. The field notes are traditionally bound and made available for research in the museum's library. Through the MVZ CLIR grant, I had the responsibility of formalizing the MVZ's archival repository and implementing a program to identify, process/describe/catalog, integrate, and make available the field notes and the many other archival collections housed at the MVZ.

I quickly learned that traditional archival best practices in access and description were neither sufficient nor appropriate in the natural history museum context. Although archives professionals are critical to the curation of historical materials, there are some nuances in the needs of natural history museums. Scientific data do not fit categorically into the traditional archival concepts of active and inactive records. Scientific data are treated and accessed much like active records, and this presents many challenges with security and preservation.

Additionally, traditional specimen cataloging is both granular and dynamic with structured relationships across individual objects. This is a departure from the current shift in archiving toward the approaches described in Mark Greene and Dennis Meissner's 2005 groundbreaking paper "More Product, Less Process." The MVZ Archives address these inherent conflicts while also using traditional archival description and dissemination. The result is mutually beneficial. Reference requests have increased exponentially since the MVZ Archives began publishing finding aids in the Online Archive of California (OAC), and our finding aids are more robust and dynamic to fit the needs of researchers.

To achieve our goals in the CLIR grant, the MVZ Archives executed the following strategy:

1. Survey all museum materials not actively cataloged by staff curators. Converse with personnel about papers that they may have inherited from former staff and faculty.
2. Install and implement Archivists' Toolkit.
3. Prioritize collections for processing and preservation.
4. Create collection records for all identified archival collections-collection building.
5. Process. Process. Process.
6. Establish portals and disseminate information.

As we worked, we developed administrative documentation surrounding archival policies. These include a mission statement, collection policy, use policy, image use and permission forms, accession forms, processing manual, imaging protocols, and numerous other administrative resources.

Initially, the materials were housed throughout the museum. Personal papers were traditionally passed down to incoming faculty and staff. Field notebooks were stored in three different locations, depending on their binding status. After an extensive survey of the museum, the Archives staff began building the collection and bringing the materials together. We identified approximately 600 linear feet of archival materials. The museum had once held a large reprint collection in a dedicated room, and the Archives staff processed the reprint collection and made room for the incoming archival collections. The Reprint Room is now a dedicated room in the Archives.

Once we established the foundation for the Archives, we explored how best to integrate archival description with specimen data. The MVZ catalogs its specimens, observations, and other biodiversity records in Arctos, a multi-institutional collection management system for natural science collections. This collaborative solution contains more than 3 million natural history museum records. In close collaboration with our staff curators, we developed a process by which to catalog localities and specimen records, and create external relationships to Arctos. I detail this process later in this paper.

The MVZ Archives has published 42 finding aids on the OAC, 35 of which are fully processed collections with encoded archive description (EAD) finding aids. Although this is only 10 percent of the collections, we anticipate that we will publish collection MARC records for the remaining 90 percent of collections by the end of the grant period.

We also provide numerous work opportunities to students. We offer internships for processing collections to graduate students attending the San Jose School of Information. We also offer exposure to archival practices, including rehousing, indexing, exhibit planning and development, and processing of small collections to our university's Undergraduate Research Apprentice Program; these students write about their experiences on the MVZ Archives Blog.

As we process collections and publish finding aids to the OAC, our reference questions are increasing exponentially. In the first year of our CLIR grant, we documented 8 reference requests; last year, there were 31, a 300 percent increase. Six months into the new fiscal year, we had 24 reference requests. Our reference questions vary across disciplines. In addition to the traditional users from the biological sciences, there are a growing number of requests from those in the humanities.

The CLIR grant allowed the MVZ to establish intellectual control over its valued archival collections and develop innovative approaches for integrating the archival collection with its specimen collection while facilitating discovery and access. We are motivated to promote and create opportunities for interdisciplinary research. We recently hired John Deck, a programmer whose previous projects include the Moorea Biocode Project, Biological Science Collection (BiSciCol) Tracker tools, and BerkeleyMapper. John's experience in informatics standards and the semantic web to integrate metadata across domain-specific databases and data workflows is ideal for creating a robust method for the online search and delivery of archival content that will be the foundation for new methods for archival data visualizations and discovery.

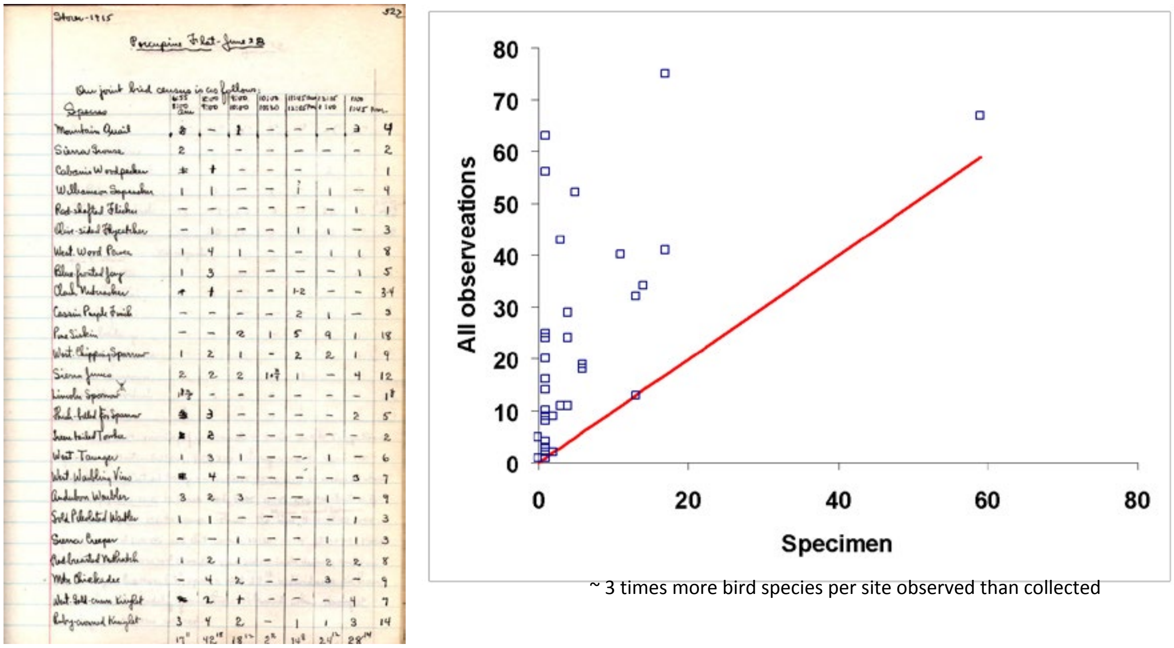
Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material? What are the issues involved in digitization (including crowd sourcing facilitated by digital versions of the resources)?

There is an urgent need to make the data contained within field notes available for assessing the impacts of rapid and ongoing environmental change. As more MVZ field notes collections are processed, researchers from across California and the country are requesting access to digitized copies. It is somewhat challenging to meet this need with just under 50 percent of our field notes scanned. Each volume contains an average of 200 pages, and as we noted earlier, there are approximately 700 bound volumes.

Furthermore, field notes contain critical observational data that our scientists can quantify and catalog. Typically, the number of observations recorded in field notes at a given locality is larger than the number of specimens collected and cataloged in Arctos. Quantifying observational data and creating observational records in Arctos would give researchers a more comprehensive understanding of changes in faunal conditions across the twentieth century.

In the example shown on the next page, three times as many bird observations are made in the field notes as there are specimens for the same location.

Field notes are handwritten and are not generally good candidates for current OCR (optical character recognition) tools. However, there is a strong citizen science community that could be engaged for transcribing field notes. With transcriptions, MVZ personnel would be able to extract data; catalog them; and relate them to specimens, localities, and other primary sources. We are continuing to explore and test different mechanisms for field note transcription.



Storer-1915 522

Porcupine Flat - June 28

Our joint bird census is as follows:

Species	6:55 8:00 am	8:00 9:00	9:00 10:00	10:00 11:30	11:45 am 12:05 pm	2:15 4:00	4:00 5:45 pm.	
Mountain Quail	[illegible]	–	1	–	–	–	3	4
Sierra Grouse	2	–	–	–	–	–	–	2
Cabanis Woodpecker	[illegible]	+	–	–	–			1
Williamson Sapsucker	1	1	–	–	1	1	–	4
Red-shafted Flicker	–	–	–	–	–	–	1	1
Olive-sided Flycatcher	–	1	–	–	1	1	–	3
West. Wood Pewee	1	4	1	–	–	1	1	8
Blue-fronted Jay	1	3	–	–	–	–	1	5
Clark Nutcracker	[illegible]	+	–	–	1-2	–	–	3-4
Cassin Purple Finch	–	–	–	–	2	1	–	3
Pine Siskin	–	–	2	1	5	9	1	18
West. Chipping Sparrow	1	2	1	–	2	2	1	9
Sierra Junco	2	2	2	1+[illegible]	1	–	4	12
Lincoln Sparrow X	[illegible]	–	–	–	–	–	–	[illegible]
Thick-billed Fox Sparrow	[illegible]	3	–	–	–	–	2	5
Green-tailed Towhee	[illegible]	2	–	–	–	–	–	2
West. Tanager	1	3	1	–	–	1	–	6
West. Warbling Vireo	[illegible]	4	–	–	–	–	3	7
Audubon Warbler	3	2	3	–	–	1	–	9
Gold. Pileolated Warbler	1	1	–	–	–	–	1	3
Sierra Creeper	–	–	1	–	–	1	1	3
Red-breasted Nuthatch	1	2	1	–	–	2	2	8
Mtn. Chickadee	–	4	2	–	–	3	–	9
West. Gold-crown Kinglet	[illegible]	2	+	–	–	–	4	7
Ruby-crowned Kinglet	3	4	2	–	1	1	3	14
	17^{11}	42^{15}	18^{12}	2^{2}	14^{8}	24^{12}	28^{14}	

~ 3 times more bird species per site observed than collected

Fig. 1: Comparison of observed bird species as recorded in field notes (left) with number of specimens collected.

Why is integrated access to data and images across scientific disciplines, institutions, and type of resource (e.g., via aggregators like iDigBio and the Biodiversity Heritage Library) important? How can this be accomplished?

Each specimen record in Arctos contains metadata describing the collecting event of the specimen, including who collected it, when, and where. The field notes contain the same data, but they include more context for the conditions of the collecting event.

Because MVZ researchers use the Grinnell method of taking field notes, the MVZ field note volumes are structured in a predictable and consistent way. Field notes typically consist of three sections: specimen catalog, journal, and species accounts. Each section has its own specified format, including placement of names, localities, and dates. Because the materials are consistent in their formats, we could reliably expect to extract data surrounding collecting events:

- *Localities.* Arctos has its own geographic name structure. The MVZ staff use this structure as the name authority, which will support future georeferencing of the materials.
- *Specimens.* We use a Related Materials Note in our finding aids to connect researchers to specimens described in a given field notebook section.

These connections allow researchers to view the inherent relationships between the field notes data and the specimens.

We are also moving toward making these connections within Arctos. This would expose the archival collections to three distributed data networks for vertebrates: MaNIS, HerpNet, and ORNIS. It would also make the materials available to the Global Biodiversity Information Facility (GBIF).

Digital management and preservation of this work, the digital data, and images: What is the current status of digital data management and preservation in your institution?

Unquestionably, digital data management and preservation are major challenges facing researchers and archivists alike. In addition to an internal image vault, the MVZ stores its digitized archival content offsite. The Texas Advanced Computing Center (TACC) provides storage of our digitized archival materials and audio, including all TIFF preservation files and all access copies.

The MVZ Archives is experiencing an influx of deposited floppy disks with data created on software that is now obsolete. And there is growing concern in general over the management of data in the scientific community. The National Science Foundation now requires that all grant proposals provide data management plans that include outlines for sharing and archiving data. Specifically, plans address file formats for long-term access, documentation for data interpretation, and copyright status. Although software and file type obsolescence is happening, data management is a proactive approach to address these problems. This is an area where archives can lend best practices in preserving data integrity, storage, and preservation. At the University of California, Berkeley, the Bancroft Library is presently implementing a digital forensics system to secure born-digital collections. This program is in its infancy, but we expect to continue exploring and improving mechanisms for digital preservation and access to born-digital collections.

▶ Barbara Mathé, Museum Archivist, American Museum of Natural History

Project titles: For the People, for Education, for Science: the American Museum of Natural History Archives (2010); and Hidden Connections: Linking Museum Expeditions, Scientists and Collections at the American Museum of Natural History (2012)

Access (resource discovery): Why and how have you made it possible for your archival collections to be found by those looking for them (or perhaps by those who did not know they existed)?

Historically and to this day, fieldwork is the source of natural science research collections. After first surveying and creating records for the scientific archives in the American Museum of Natural History (AMNH) with funding from CLIR and the Institute of Museum and Library Services (IMLS), we are using our second Hidden Collections grant to focus on AMNH archives from museum expeditions. We want to better document and relate collecting events, noting particularly the people who were on the expeditions at the same time, place, and circumstances when the collections were made. As Grinnell pointed out, this information is essential to the full significance of the specimen collected, whether for the systematic study of species or for use in ecological research. AMNH scientists have not used Grinnell's data fields on a consistent basis, although currently there is a growing interest in doing so. For example, a scientist returning from Papua, New Guinea, on a recent Explore 21 field trip was just extolling the virtues of Grinnell-formatted field notes.

Founded in 1869, the AMNH will soon celebrate its 150th year. I am the museum archivist, based in the research library and charged with coordinating efforts with others in the museum who manage archival collections. I report to the library director, who reports to the provost, who is senior vice president for science. There are five divisions in the administrative unit of Science, of which the library is also an administrative unit. The divisions are Anthropology, Physical Sciences, Invertebrate Zoology, Vertebrate Zoology, and Paleontology (both vertebrate and invertebrate). The Division of Anthropology has a professional archivist, and her time is split between archives and managing traveling loans. Other divisions either have retired scientific assistants or curatorial associates, whose chief responsibility is the management of scientific collections, overseeing their archives.

The specimens and objects in the scientific collections now number more than 33 million. We can estimate 25,000 linear feet of archives across the museum, in the science departments, and in the research library. This estimate was the result of the timely confluence of our first CLIR grant in 2010 to create minimal catalog records for the archives in the library, with an IMLS risk assessment grant that allowed us to survey the archives in the science departments. These two projects resulted in the creation of some 3,275 minimal-level catalog records for the archival collections in AMNH Science, 1,400 in the library and 1,875 in the science divisions. We call these minimal records "skinnies." Although they are mapped to MARC and to EAD, and published online so that people might discover their existence, they are hardly full finding aids. Our first CLIR grant also resulted in 21 fully descriptive finding aids.

Many of these archival collections, especially the field books and notes, relate directly to the scientific collections. This is a gold mine of related data of all kinds, but the knowledge of the existence of, and particularly the relationships across, the materials is still largely dependent upon the accumulated institutional knowledge of the staff. One key to integrating access across archives and scientific collections, that are described only minimally might be found in the relationships between expeditions and the many scientists who traveled on them. For example, a researcher looking at the circumstances surrounding the collection of a bird could find that a herpetologist on the same expedition had documented the environment at the same place and time the bird was collected. With a minimal EAD collection record of that herpetologist's field book, listing only her name, the information would be hidden. But if a record for the expedition listed both the name of the herpetologist and the name of the ornithologist who collected the bird, each assigned an individual identifier and related by the identifier of the expedition, then the connection can be made. With such an enormous collection, approaching access through minimal records that can be linked and then enhanced over time seems the best approach. We just cannot describe every record one by one and expect to get results in a reasonable period of time. When we began the second CLIR project, we anticipated the development of linked data applications and are looking at a number that have recently been launched.[1]

Using Encoded Archival Context–Corporate Bodies, Persons, and Families (EAC–CPF) to create records that can normalize the names of the expeditions and those associated with them will allow us to integrate access to collections (whether archival or specimen), even if they are described on a minimal level. We came to this approach because an earlier library project had

[1] As of early November 2015, the AMNH team will be working to develop a proof of concept for linking data across different collections across the scientific and archival collections in the museum with a triplestore.

documented close to 1,000 museum expeditions. We updated that work with the information found in vertical files for expeditions and for people associated with the museum. With the help of many volunteers over the years, the library had created spreadsheets that captured basic data. Happily, the data fields were a close match to fields in the EAC–CPF data standard, and we set out to verify and use this data as a starting point for the project. Incidentally, we have similar data for museum departments, halls, and temporary exhibitions.

So we embarked, once again, on our effort to create minimal level records, this time for entities, specifically AMNH expeditions and personal names. We have also created templates and protocols for our interns to use to create fully developed EAC–CPF records, recognizing that the biographical or historical note for the entity can then be repurposed for any collection that has an EAD record, using the EAC–CPF biographical/historical note related collections. For example, there are five archival collections from the Whitney South Sea Expedition, one in the library, one in the Department of Herpetology, and three in the Department of Ornithology. The EAC–CPF record for the entity, "Whitney South Sea Expedition," is being repurposed for the EAD finding aid records for all five collections.

The result is a large number of minimal-level EAD and EAC records to manage, as well as a number of records more fully describing collections and entities. Anticipating this result, we committed to develop a cyberinfrastructure for our archival records in our CLIR Hidden Collections grant proposal in 2012.

The team includes Becca Morgan, project archivist, and Iris Lee, metadata analyst (both veterans of the first CLIR/IMLS project), along with Nick Krabbenhoeft and Bill Levay, who have added their technical insight and expertise as interns and subsequent volunteers. Neither CLIR project would have been possible without our interns and the many volunteers who are listed on our Hidden Collections blog.

After developing our functional requirements for managing the records, we found that there is no archival management system that fulfills our requirements for EAC records, so we are using xEAC. For EAD records, it was a close contest between ArchivesSpace and Atom, but the decision was made to use ArchivesSpace (Lee 2015).

In response to the extensive work we did developing EAC–CPF records, as well as our process methodology and documentation during this project, we were invited to join SNAC (Social Networks and Archival Context), a collaboration between the U.S. National Archives and Records Administration, the Institute for Advanced Technology in the Humanities at the University of Virginia, the California Digital Library, and the School for Information Science at the University of California, Berkeley. According to the SNAC home page, the project "demonstrates the feasibility of separating the description of persons, families, and organizations—including their socio-historical contexts—from the description of the historical resources that are the primary evidence of their lives and work."

We also just received funding from the Leon Levy Foundation to begin cataloging the 3,224 AMNH Field Books identified across the museum. Before the advent of our funding from CLIR and the IMLS, we would not have known how many we have or where they might be found.

Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material? What are the issues involved in digitization (including crowd sourcing facilitated by digital versions of the resources)?

Digitization is the next step in establishing wider integrated access to museum collections of all kinds. Because of the scale of our collections, we are releasing minimal records. If we can attach scans to those records, we will be in a position to engage the citizen scientists in crowd sourcing the transcription of these materials, revealing more detailed information that can be added (and linked to and from) those records.

Not only can making scanned materials available to researchers allow them to work remotely, but also they will not have to wait for collections to be fully processed before they begin their research. Obviously, there are privacy and other issues relating to the content of the collections, but archivists will learn how to manage them.

Why is integrated access to data and images across scientific disciplines, institutions, and type of resource (e.g., via aggregators like iDigBio and the Biodiversity Heritage Library) important? How can this be accomplished?

Natural science collections span libraries, archives, and museums. A species is "named" upon the first publication of its description by the person who found it. For this reason, rare books are active documents in natural science research. The Biodiversity Heritage Library is a consortium of libraries that are digitizing pre-1923 natural science publications. They are also beginning to include digitized field books. It is not unreasonable to foresee integrated access in aggregators like the Biodiversity Heritage Library and iDigBio, as well as other aggregators like VertNet and MorphoBank, as a long-term goal to make it possible for researchers to seek information across institutions, whether libraries, archives or museums. And by using globally unique identifiers and linked data applications, it may soon be possible to search across the aggregators.

Digital management and preservation of this work, the digital data and images: What is the current status of digital data management and preservation in your institution?

The AMNH is fortunate to be hosting a National Digital Stewardship Resident who has been interviewing the museum's scientific curators and staff about the state of digital preservation and management of their scientific research data. The data include enormous files, such as genomic sequences; three-dimensional CT scans of specimens; and high-definition film as well as collection data. Her report will form the basis of our effort to address the very complex solutions involved in establishing and staffing a permanent program for digital preservation and management at the AMNH. In time, we hope it will result in a trustworthy digital repository for the museum's scientific research and collection data in digital form.

Addendum: As a result of the ongoing work done by on the CLIR projects, followed by that done by the AMNH NDSR resident, we are beginning to move onto the next phase of our work to produce a cyberinfrastructure at the AMNH. A poster that we designed and presented at the recent Hydra Connect meeting illustrates a conceptual model of our current information landscape as compared to an sketch of how we imagine organizing our assets into a cyber-infrastructure. The model is based upon the OAIS model and can be seen on the process tab on our AMNH Hidden Collections site: http://images.library.amnh.org/hiddencollections/process/

▶ Russell D. “Tim” White, Director of Collections and Operations, Peabody Museum of Natural History, Yale University

Project title: From DNA to Dinosaurs: The Globalization of Science in America and the Development of a University Natural History Museum (2010)

Access (resource discovery): Why and how have you made it possible for your archival collections to be found by those looking for them (or perhaps by those who did not know they existed)?

Through seven generations of faculty and students at Yale University, the Peabody Museum of Natural History has acquired a treasure trove of historical and scientific documentation that relates to some of the earliest scientists in America (e.g., Benjamin Silliman) and some of the earliest organized (and unorganized) natural history collections. The ability to keep these materials in an organized, accessible, and retrievable fashion has bolstered the value and use of Peabody’s specimens and artifacts, and it has led to many new discoveries in the collections and in the field. Moreover, it is important in documenting the history of scientific inquiry.

At the Peabody, archival materials from previous faculty and students have been “managed” and housed within the 10 scientific collection divisions. In some divisions, archival collections are well organized, properly housed in archival quality materials, and cataloged; in others, collections have been at best physically located in a designated area with little or no arrangement. While faculty members are active, their field notes, maps, etc. remain in their possession, generally kept in their offices and labs, and are only turned over to the museum after retirement.

For too long, archival collections at the Peabody have been seen as supporting documentation for specimens and artifacts; they have not been recognized as significant historical documentation on their own. However, field notes, maps, photographs, diaries, and other materials often document the details of a collecting event and scientific discovery. As part of the CLIR project, archival collections within each curatorial division were assembled and assessed for preservation; diversity of materials; relationship to Yale faculty and students; and documentation of specimens, artifact discovery, and use. Access to the cataloged archives and special collections of the Peabody Museum has opened up new avenues for research beyond the bounds of natural science, including the history of early science in America and the exploration of the American West.

Our CLIR-funded project, From DNA to Dinosaurs, allowed us to assess and inventory all of our archival holdings in the 10 curatorial units plus the museum archives, appropriately rehouse materials, and catalog our collections using the EAD standard in a manner that allows for maximum accessibility. In fall 2014, the Peabody launched 78 finding aids that are available on the Yale Finding Aid Database, and another 20 are in production.

With collections from A to Z at the Peabody Museum, including anthropology, botany, geology, paleontology, and zoology, the discipline-specific historical practices need to be evaluated for methods of documentation of the collecting event, the geospatial occurrence, and the geologic age. By using original documentation, we have developed workflows in the curatorial divisions to catalog our specimen and artifact collections to better define best practices for this type of collection, which has led to improved cataloging and informatics, as well as increased productivity.

In 2016, the Peabody will be celebrating the 150th anniversary of its establishment as a Yale University museum with a gift from George Peabody. One of the celebratory events is a book by Richard Conniff, who is writing about the 15 faculty curators and significant events in our history. Use of the Peabody finding aids and unfettered access to the well-organized archives has made discovery for this project feasible and possible—a situation that did not exist prior to this Hidden Collections project.

Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material? What are the issues involved in digitization (including crowd sourcing facilitated by digital versions of the resources)?

Access to digitized versions of Peabody's archival materials enhances our ability to develop methods for better management and dissemination of specimen-level information. Also, it offers a better understanding of the people involved in these collections. The breadth of work of Benjamin Silliman, Benjamin Silliman, Jr., and James Dana in the early nineteenth century helped to define Yale's nascent programs in science education for undergraduate students. Their archival materials provide insights into the development of a teaching program for science at Yale and other insights into early geological investigations and classifications, such as *Dana's System of Mineralogy* (first published in 1837). In the late nineteenth century, O. C. Marsh and A. E. Verrill defined advance training for biological and paleontological students, and the establishment of Yale as a leading center for paleontology and marine biology in North America (Conniff, in press).

Why is integrated access to data and images across scientific disciplines, institutions, and type of resource (e.g., via aggregators like iDigBio and the Biodiversity Library) important? How can this be accomplished?

Within the natural history community, there are well-established tools for disseminating specimen and collecting event information through the Global Biodiversity Information Facility, iDigBio, VertNet, and other aggregators that specialize in collating morphological information through images (e.g., Morphbank, MorphoBank). Through the efforts of the Biodiversity Heritage Library, some specimen- and collection-level documentation is available, but to date, there is no one aggregator that pulls information from specimen, collecting events, and historical documentation into a central repository.

Some institutions have developed search interfaces to look across collections from disparate sources. At Yale University, the Cross Collections Discovery (CCD) aggregates materials from the Peabody Museum, Yale art galleries, and libraries in one search portal, allowing for discovery of vastly different material in one location.

Digital management and preservation of this work, the digital data and images: What is the current status of digital data management and preservation in your institution?

At the Peabody Museum, archives and special collections are cataloged using KE EMu, the same database management application that is used for cataloging and tracking Peabody's specimen and collecting event information. Specimen information is distributed to numerous biodiversity portals, and archival information is published through the Yale Finding Aid Database. With both types of collections being cataloged using KE EMu, we can cross-reference different types of associated information in a consistent and structured manner. Over the past three years, the Peabody has digitized more than 1,000 cataloged field notes, ledgers, and other forms of original documentation using a robotic book scanner and linked these digitized documents to specimens and collecting events in KE EMu.

One challenge for the Peabody is to disseminate these linked data sets of archival information and specimen or artifact information to the broader community. Another challenge is to extract information from these types of digitized legacy documents (e.g., institutional catalogs) and original documentation (e.g., field notes). Transcribing targeted information such as people's names, geographic places, taxonomic names, geological ages and rock units, and dates is a goal. Crowd sourcing this kind of activity also seems highly desirable.

▶ Rusty Russell, Program Director for Collections and Informatics, United States National Herbarium, Smithsonian Institution

Project title: Exposing Biodiversity Fieldbooks and Original Expedition Journals at the Smithsonian Institution (aka The Field Book Project) (2009)

Access (resource discovery): Why and how have you made it possible for your archival collections to be found by those looking for them (or perhaps by those who did not know they existed)?

This question gets at what was precisely the inspiration for The Field Book Project (FBP). My frustrating attempt to locate field notes from the U.S. Exploring Expedition (1838–1842) led me to more than a dozen institutions. Although there were some electronic services that provided information about these materials, most content was found through inferential searching or simple guesswork. And the record is still incomplete. Similarly, over many years, I have received innumerable inquiries from colleagues looking for original sources (primarily field notes) of information documenting botanists' field activities. Sometimes they knew we had it, sometimes they thought we might have it, and sometimes they were simply casting a wide net. In hindsight, the idea that we should address field books in the same way that we catalog books or inventory natural history specimens seems painfully obvious. And yet, except for materials that were deposited in archival repositories, most of these items were "hidden" because of the lack of a local archives; libraries' inclinations not to claim them; or a well-intentioned, but ill-advised, tradition of biologists "caring" for their predecessor's scholarly work in support of biodiversity research.

My co-principal investigator Anne Van Camp, director of the Smithsonian Institution Archives (SIA), had also identified the need for better descriptions of field books. The SIA contains thousands of field books but, as is the standard in archives, descriptions of individual items within field book collections were minimal. Recognizing our mutual goal of improving access to field books, we envisioned a Field Book Registry that would serve as a global source for locating and describing information within field notes and other field research materials.

Biodiversity research begins with field books, primary source documents that record information about what was observed, discovered, and collected in nature. Because of their integral relationship to specimen collections, field books are often consulted by researchers for a variety of scholarly inquiries. Yet it is their close relationship to the specimen collections, as well as their nature as primary source documents, that makes their categorization as objects ambiguous. Biologists frequently consult field books when they are reconstructing field activities that resulted in collected specimens. Are field books museum objects that should physically reside close to specimen collections and be described and made available in a similar way? Are they archival materials that should be stored and preserved in archives and described in finding aids? Or should these book-like resources be considered library objects to be cataloged as individual volumes in an online catalog? Perhaps all of these aspects are valid.

The Field Book Project began as a joint initiative of the National Museum of Natural History (NMNH), and the SIA. Together, they applied for and received funding from CLIR to uncover the hidden collections of field books at the Smithsonian. The early years of the Field Book Project consisted of two phases. During the initial phase, the Field Book Project focused on locating field notes throughout the Smithsonian's many research facilities and creating catalog records for these field books within a local database. Phase two saw these records being migrated to a robust, XML-driven pilot registry that will be opened to other institutions so that they can add catalog records for field books in their repositories.

Field books at the Smithsonian Institution are maintained by multiple departments or units: the NMNH science departments, the Smithsonian Institution Libraries, the National Zoo, the Smithsonian Tropical Research Institute (STRI), and the SIA. Some departments, like the Department of Botany, had previously created inventory lists with basic information to help staff members and patrons find the field books they needed. In other NMNH departments, the SIA had created basic finding aids in the 1970s and 1980s, and had done some light processing, but the collections remained in the custody of their respective departments. Earlier SIA finding aids had been very helpful as a foundation for description, but are now outdated in many cases. Field books are held in the stewardship of museums, archives, and libraries, and therefore benefit from a flexible, yet consistent, standards that combine descriptive practices from all three fields. Our project has drawn from metadata and encoding schemas and content standards across the disciplines to create a hybrid descriptive framework that bridges collection- and item-level description.

Collection-level metadata is a hallmark of archival description. Archival finding aids describe materials as collections, rarely becoming more granular than brief folder-level descriptions. In archival description, providing the context in which materials were created is as important as describing the materials themselves. Unlike published works, which are self-justifying and stand-alone objects, archival documents are like pieces of a puzzle; although they are useful on their own, they need other documents within the collection to provide context and tell a full story.

Libraries have long been innovators for item-level metadata, which have traditionally emphasized access points such as authors or creators, subjects (e.g., topics, locations, names), and titles. The access points of geographic and topical coverage are incredibly important for addressing the bulk of known research needs for primary source field notes. Such access points allow researchers to more easily pinpoint desired volumes. Prolific scientists may create more than 100 field books over the course of their career, spanning multiple collecting events across the globe. Distinguishing one volume from the next based on content, therefore, becomes important for meeting information needs. And because libraries are precedent setters for controlled vocabulary within access points, we follow library descriptive practices for cataloging at the item level.

We define a "collection" as any group of field books with a unifying relationship. Field book collections can be assembled in many ways; our collections, however, are usually grouped by collector or expedition. For example, a collection grouped by the collector Alexander Wetmore would consist of field books created or owned by Wetmore. Alternatively, a collection grouped by the "United States Exploring Expedition" might consist of field books created by various individuals who participated in that expedition. Less frequently, our collections are assembled by an organization as a creator. However they are grouped, the way the field books were physically organized, with respect to the provenance and order in which they were received and maintained prior to our involvement, determines the collections, in accordance with archival practice.

Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material? What are the issues involved in digitization (including crowd sourcing facilitated by digital versions of the resources)?

The initial CLIR grant provided generous funding for our cataloging efforts in the Field Book Project. Early in our workflow designs, we recognized that many field books needed remedial conservation to survive the necessary handling required of this effort. Conservation tasks needed to be a concurrent feature in the workflow with discovery and cataloging. We leveraged the CLIR grant to obtain funding from the National Park Service's Save America's Treasures program to support our conservation efforts. This allowed us to stabilize objects for cataloging and for future digital imaging. In addition, the Smithsonian Women's Committee awarded funds to continue conservation and begin the task of digitally imaging field books at the page level.

What does it mean to digitize something? Technically, it is simply converting data from analog to digital form. Digitization can include producing a digital image of a field book page. It also describes the task of converting (transcribing) text from specimen labels, correspondence, or pages in a field book. However, the term *digitization* has become synonymous with digital imaging, and the term *transcription* is now used to describe the digitization of text. For the purposes of this account, I will refer to digital imaging and transcription.

Digital imaging does not improve access to content. Digital indices, keywords, and web linkages provide that service. That is why it was important to design a catalog record for field books that provided enough metadata to facilitate the process of searching for field book content. The catalog records that we created employ multiple existing standards for managing objects; collections of objects; and inherent connections to people, institutions, and defined collecting events. Once found, a digital image can provide a wealth of information, either explicit or interpreted. In the case of field book pages, it also provides an unambiguous account of a field collector's activities and experiences. It is unambiguous because it is original. There has been no modification, transcription, or transfer of information.

Digital imaging at the page level was a critical piece of our vision for the Field Book Project. Therefore, the design of the Field Book Registry was originally seen as supporting page-level navigation of digitally imaged field books. More importantly, page-level imaging was understood to be the necessary bridge to fully word-searchable original source materials. Only when every word in a field book is essentially an index term can the incredible, yet latent, power of field books be fully realized. For example, we could deliver every mention of a native use of *Artocarpus altilis* from Samoa prior to 1950.

The task of transcribing every word on every page in every field book, however, will be the most complicated and time-consuming step toward this goal. Currently, the Field Book Project is using the Smithsonian Institution Transcription Center and its army of volunteers to perform transcription services. Our project competes with other Smithsonian Institution projects for the attention of transcribers, and progress is slow. Other transcription services must be considered if the field books are going to become a critical and productive resource for biodiversity research.

Why is integrated access to data and images across scientific disciplines, institutions, and type of resource (e.g., via aggregators like iDigBio and the Biodiversity Heritage Library) important? How can this be accomplished?

Although the need to aggregate related content across institutions, people, and events is what inspired the Field Book Project, the Holy Grail is being able to recognize and deliver inherent relationships between field books, specimens, and published citations, or what is now routinely called "connecting content." Aggregators of institutional content, disciplinary content, and object types serve an important role in our need to see and interpret data as never before. But although these efforts have been highly productive, we have only scratched the surface in our plan to marshal technology and informatics toward seamless access to related content.

Recently, the Field Book Project has joined forces with the Biodiversity Heritage Library in a collaboration that will combine BHL content with field books from more than a dozen major natural history partners.

Digital management and preservation of this work, the digital data and images: What is the current status of digital data management and preservation in your institution?

The Field Book Project is an unusual example of cross-bureau collaboration within the Smithsonian (i.e., the National Museum of Natural History and the SIA). From the beginning, the Office of the Chief Information Officer was engaged to help oversee the technical development and to manage resulting content.

References

Conniff, Richard. In press. *House of Lost Worlds: Dinosaurs, Dynasties, and the Story of Life on Earth.* Yale University Press.

Gaines, Richard V., H. Catherine W. Skinner, Eugene E. Foord, Brian Mason, and Abraham Rosenzweig. 1997. *Dana's New Mineralogy,* 8th ed. New York: John Wiley & Sons, Inc.

Greene, Mark A. and Dennis Meissner. 2005. More Product, Less Process: Revamping Traditional Archival Processing. *The American Archives* 68 (Fall/Winter): 208-263.

Grinnell, J. 1910. The Methods and Uses of a Research Museum. *Popular Science Monthly* 77 (January 1): 163–169.

Lee, Iris. 2015. Making Decisions…Archives Space or AtoM? Blog post, July 27, 2015, in "Hidden Collections: Stories From the Archive." Available at http://images.library.amnh.org/hiddencollections/2015/07/making-decisions/#more-4847

MVZ Archives Blog: https://mvzarchives.wordpress.com/

SNAC (Social Networks and Archival Context) website: http://socialarchive.iath.virginia.edu/