Grinnell to GUIDs: Connecting Natural Science Archives and Specimens

"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." From: "The Methods and Uses of a Research Museum" by Joseph Grinnell (1915), *Popular Science Monthly* 77: 163–169.

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, and observations that did not result in specimens collected, interactions with local or indigenous populations, and other data, e.g. 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, reflecting a growing momentum toward associating access to all types of resources related to natural science regardless of type of resource or whether they are held in archives, museums or libraries. While 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—whether held in museums, archives or as publications in libraries. As the development of linked data applications using globally unique identifiers (GUIDs) proceeds, this goal can become a reality.

Four recipients of CLIR Hidden Collections grants in natural science museums, two archivists and two scientists, discuss questions regarding how they view and approach access for resource discovery, the effects of digitization, integrated access in aggregators and the issues of long term digital preservation.

CHRISTINA FIDLER, MUSEUM ARCHIVIST
MUSEUM OF VERTEBRATE ZOOLOGY AT THE UNIVERSITY OF CALIFORNIA BERKELEY
Cataloging Hidden Archives of the Museum of Vertebrate Zoology: Increasing Integration and Accessibility for Interdisciplinary Research (2011)

QUESTION 1—Access (resource discovery): Why and how we each approached making it possible for our archival collections to be found by those looking for them (or perhaps by those who didn't know they existed).

Beginning in 2003, the Museum of Vertebrate Zoology 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. In order to conduct this extensive resurvey, researchers relied primarily on the archival field notes which document the floral and faunal conditions of the same transects roughly 100 years ago. It is an astonishing testament to the importance of these historical data. The resurvey was a catalyst for the Museum's CLIR grant, that 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 year, 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"[1]

As mentioned above, Joseph Grinnell developed a means for capturing data while collecting and observing in the field. The Grinnell method of taking field notes is still practiced much in its original form at the MVZ. This has resulted in over 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 was given 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. While archival professionals are critical to curating historic materials, there are

[1] Grinnell, J. (January 01, 1910). The methods and uses of a research museum. Popular Science Monthly, 163-169.

some nuances in approaching the need of natural history museums. Scientific data does not fit categorically in the traditional archival concepts of active and inactive records. Scientific data is 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 normalized relationships across individual objects. This is a departure from the current shift in archives towards the approaches described in Mark Greene and Dennis Meissner's 2005 ground-breaking paper "More Product, Less Process". The MVZ Archives addresses these inherent conflicts while also utilizing traditional archival description and dissemination. The result has been mutually beneficial. Reference requests have increased exponentially since the MVZ Archives began publishing finding aids on the Online Archive of California (OAC) and our finding aids are more robust and dynamic to fit the needs of our 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 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. Portals and dissemination.

Throughout this timeline, 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 began collection building and bringing the materials together. We identified ~600 linear feet. The museum once held a large reprint collection in a dedicated room. The Archives processed the Reprint collection and made room for the incoming archival collections. The Reprint Room is now a dedicated Archives Room.

Once we established the foundation for the Archives, we could begin to explore how to better 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 serves over 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 records on the Online Archive of California (OAC), 35 of which are fully processed collections with EAD encoded finding aids. While this is only 10% of the collections, we anticipate that we will publish collection stub MARC records for the remaining 90% 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 UC Berkeley's Undergraduate Research Apprentice Program (URAP). Our URAP students write about their experiences on the the MVZ Archives Blog (https://mvzarchives.wordpress.com/).

As we process collections and publish finding aids to the OAC, our reference questions are increasing exponentially. In our first year of the CLIR grant, we documented eight reference requests. We documented 31 reference requests last year. This is a 300% increase. Six months into the new fiscal year, we have recorded 24 reference requests. Our reference questions vary across disciplines including traditional users in the biological sciences and a growing increase in the humanities.

The CLIR grant allowed the MVZ to develop intellectual control over its valued archival collections and to develop innovative approaches for integrating the archival collection with its specimen collection while facilitating discovery and access. We have been 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 online search and delivery of archival content that will be the foundation for new methods for archival data visualizations and discovery.

QUESTION 2—Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material and the issues involved in digitization (including crowd sourcing facilitated by digital versions of the resources.)

Presently, there is an urgent need to make the data contained within field notes available for assessing impacts of rapid and ongoing environmental change. As more MVZ field note 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% of our field notes scanned. Each volume contains an average of 200 pages and there are ~700 bound volumes.

Furthermore, field notes contain critical observational data that our scientists can quantify and catalog. Typically, there are more observations recorded in field notes at a given locality compared with the number of specimen 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 20th century.

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

Field notes are handwritten and are not generally good candidates for current OCR 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 these 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.

QUESTION 3—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 (BHL) 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 and 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 Archives uses this structure as its name authority. This 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 note data and the specimens.

We're also moving towards making these connections within Arctos. By doing so, it would expose the archival collections to three distributed data networks for vertebrates; MaNIS, HerpNet, ORNIS. It would also make the materials available to the Global Biodiversity Information Facility (GBIF).

QUESTION 4—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 MVZ is a partner with the Texas Advanced Computing Center (TACC), which provides storage of our digitized archival materials and audio. This includes all TIFF preservation files and all access copies.

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

BARBARA MATHE, MUSEUM ARCHIVIST
AMERICAN MUSEUM OF NATURAL HISTORY
For the People, for Education, for Science: the American Museum of Natural History Archives (2010)
Hidden Connections: Linking Museum Expeditions, Scientists and Collections at the American Museum of Natural History (2012)

QUESTION 1—Access (resource discovery): Why and how we each approached making it possible for our archival collections to be found by those looking for them (or perhaps by those who didn't 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 with funding from CLIR and IMLS, our second Hidden Collections grant is focusing on AMNH archives from Museum expeditions. We want to better document and relate collecting events: 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. It should be noted, however, that AMNH scientists did not use Grinnell's data fields on a consistent basis, although currently there is a growing interest in doing so. A scientist returning from Papua New Guinea on a recent Explore 21 field trip was just extolling the virtues of Grinnell formatted field notes.

As context and background: the AMNH was founded in 1869 and will soon celebrate its one hundred and fiftieth year. I'm 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 Science, of which the Library is also an administrative unit. The divisions are Anthropology, Earth and Planetary Sciences, Invertebrate Zoology, Vertebrate Zoology, and Paleontology (both vertebrate and invertebrate).

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 for managing scientific collections, overseeing their archives.

The specimen and objects in the scientific collections now number over 33 million. We can now estimate 25,000 linear feet of archives across the Museum, in the science departments and in the research library. This estimate of the archival collections 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 3275 *minimal level* catalog records for the archival collections in AMNH Science (1400 in the library and 1875 in the science departments.) Note: these are minimal records. We call them "skinnies." But they are mapped to MARC and to EAD and published online so that people might discover their existence but they are hardly full finding aids. Our first CLIR grant also resulted in 21 fully described 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 people working there. So one key to integrating access across archives and scientific collections, described only minimally, seemed to us to 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, the same place and time the bird was collected. With a minimal EAD collection record of that herpetologist's field book that listed 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 their own 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 can't 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 development of linked data applications and we are looking at a number that have recently been launched.

Using 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 or archival or specimen) even if they are described on a minimal level. We came to this approach because an earlier Library project had documented close to a thousand 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, this data was 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 creating minimal level records, this time for entities, specifically AMNH expeditions and personal names. We've also developed templates and protocols for our interns to use to create fully developed EAC-CPF records, recognizing that the biographical or historical note developed for the entity can then be repurposed for any collection that has a record in EAD, using the EAC-CPF biog/hist note related collections. For example: there are five archival collections from the Whitney South Sea Expedition, one in the Library, one in Herpetology and three in Ornithology. The EAC-CPF record for the entity, Whitney South Sea Expedition, is being repurposed for the EAD encoded finding aid records for all five collections.

The result is we now have a large number of minimal level EAD and EAC records to manage as well as a number of more fully described collections and entities. Anticipating this result, we committed to develop a cyber-infrastructure for our archival records in our CLIR Hidden Connection grant proposal.

The team includes Becca Morgan, project archivist, Iris Lee, Metadata Analyst (both veterans of the first CLIR/IMLS project) along with Nick Krabbenhoeft and Bill Levay, who add their technical insight and expertise as interns and subsequent volunteers. Both CLIR projects would not have been possible with our interns and many volunteers who are listed on our blog.

After developing our functional requirements for managing the records, we have found that there is no archival management system that fulfills our requirements for EAC records so we are using xEAC. For EAD records, it's a close contest between ArchiveSpace and Atom and a decision will be made by early March to be announced at the CLIR Symposium and also on our Hidden Collections blog which has considerable documentation about all aspects of what we're calling the AMNH Archive Project, well on its way to becoming our archive program.

In response to the extensive work we have done developing EAC-CPF records and process methodology and documentation during this project, we have been invited to join SNAC pilot project. SNAC hosted by the U.S. National Archives and Records Administration (NARA), is, as described on their website, "…demonstrating 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 have also just received funding from the Leon Levy Foundation to begin cataloging the 3224 AMNH Field Books identified across the Museum. Before the advent of our funding from CLIR and IMLS, we would not have known how many we have or where they might be found.

QUESTION 2—Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material and the issues involved in digitization (including crowd sourcing facilitated by digital versions of the resources.

Digitization is the next step in 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 the 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.

Making scanned materials available to researchers can allow researchers to work remotely and not have to wait for collections to be fully processed. Obviously there are privacy and other issues relating to the content of the collections but archivists will learn how to manage them.

QUESTION 3—Why is integrated access to data and images across scientific disciplines, institutions and type of resource (e.g. via aggregators like idigBio and BHL) 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. This is why 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 fieldbooks. It is not unreasonable to foresee integrated access in aggregators like the Biodiversity Heritage Library,(BHL) and iDigBio as well as other aggregators like VertNet and MorphoBank as a very long term goal allowing researchers to see across institutions, whether libraries, archives or museums for information. And by using GUIDs and linked data apps it may soon be possible to search across the aggregators.

QUESTION 4—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?

AMNH is fortunate to be hosting a National Digital Stewardship Resident who has been interviewing the Museum's 41 scientific curators and their staff about the state of digital preservation and management of their scientific research data. The data includes enormous files like genomic sequences and three-dimensional CT scans of specimens along with high definition film as well as collection data. Her report will form the basis of beginning to address the very complex solutions involved in establishing and staffing a permanent program for digital preservation and management at the AMNH which in time, we hope will result in a Trustworthy Digital Repository for the Museum's science and collections in digital form.

RUSSELL D. "TIM" WHITE
PEABODY MUSEUM OF NATURAL HISTORY, YALE UNIVERSITY
From DNA to Dinosaurs: The Globalization of Science in America and the Development of a University Natural History Museum (2010)

QUESTION 1—Access (resource discovery): Why and how we each approached making it possible for our archival collections to be found by those looking for them (or perhaps by those who didn't know they existed).

Through seven generations of faculty and students at Yale University, the Peabody Museum of Natural History has 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. Being able to relate these materials in an organized, accessible and retrievable fashion, Peabody's specimens and artifacts have bolstered their value and use and have led to many new discoveries in the collections and in the field, and are 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 ten scientific collection divisions. Whereas in some collection divisions, archival collections are well organized, properly housed in archival quality materials and cataloged while others have been at best physically located in a designated area with little or no arrangement. While a faculty member is active, their field notes, maps, etc. reside in their possession in their offices and labs, and are only turned over to the Museum after retirement.

For all too long, archival collections at the Peabody have been seen as supporting documentation for specimens and artifacts and not recognized as significant historical documentation on their own. Field notes, maps, photographs, diaries, etc. 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 Grant, *From DNA to Dinosaurs…* has allowed us to assess and inventory all of our archival holdings in the ten curatorial units plus the Museum Archives, rehouse materials in appropriate materials and catalog our collections using the EAD standard in a manner that allows for maximum accessibility. The Peabody has launched 78 finding aids that are available on the Yale Finding Aid Database last fall, and have another 20 finding aids in production.

With collections from A to Z, 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, work flows in the curatorial divisions have been developed to catalog our specimen and artifact collections using this "supporting documentation" to better define best practices for this type of collections, which has led to improved cataloging and informatics and increased productivity.

In 2016 the Peabody will be celebrating our 150th anniversary of our establishment as a Yale University museum with a gift from George Peabody. One of the celebratory events is a book by the author 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.

QUESTION 2—Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material and 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 and 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 19th century went a long way to defining Yale's nascent programs in science education for undergraduate students. Access to their archival materials offers 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 (1837). In the late 19th 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.

QUESTION 3—Why is integrated access to data and images across scientific disciplines, institutions and type of resource (e.g. via aggregators like idigBio and BHL) 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, and 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 specimens, 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, Yale art galleries and libraries in one search portal allowing for discovery of vastly different material in one location.

QUESTION 4—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, 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, the ability to cross-reference different types of associated information is achievable in a consistent and structured manner. Over the past three years, the Peabody has digitized more than 1000 cataloged field notes, ledgers and other forms of original documentation using a robotic book scanner and have linked these digitized documents to specimens and collecting events in KE EMu. One challenge for the Peabody is how we disseminate these linked data sets of archival information and specimen or artifact information to the broader community. Another challenge is how we 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. Crowdsourcing this kind of activity also seems highly desirable.

RUSTY RUSSELL, PROGRAM DIRECTOR FOR COLLECTIONS & INFORMATICS
UNITED STATES NATIONAL HERBARIUM, SMITHSONIAN INSTITUTION
Exposing Biodiversity Fieldbooks and Original Expedition Journals at the Smithsonian Institution (aka The Field Book Project) 2009 -

QUESTION 1—Access (resource discovery): Why and how we each approached making it possible for our archival collections to be found by those looking for them (or perhaps by those who didn't know they existed).

This question is attempting to get at what was precisely the inspiration for The Field Book Project (FBP). My frustrating attempt to locate field notes from the United States Exploring Expedition (1838-1842), led me to over a dozen institutions. While there were examples of electronic services that provided information about these materials, most content was found through inferential searching or simple guesswork. And the record is still incomplete. Over many years, I have received innumerable inquiries by 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 we catalog books or inventory natural history specimens seems painfully obvious. And yet, except for materials that were deposited in archival repositories, the majority of these items were "hidden" due to 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-PI Anne Van Camp, Director of the Smithsonian Institution Archives (SIA), had also identified the need for better description of field books. 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 (FBR)[i] that

would serve as a global source for location and information for field notes and other field research materials.

Biodiversity research begins with field books, primary source documents which 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 inquiry. Yet it is their close relationship to the specimen collections, as well as their nature as primary source documents that make their categorization as objects ambiguous. Field books are frequently consulted when biologists 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 these 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 aspects of all of the above are valid.

The FBP began as a joint initiative of the National Museum of Natural History (NMNH), and the Smithsonian Institution Archives (SIA). Together, NMNH and SIA applied for and received funding from the Council of Library and Information Resources (CLIR) to uncover the “hidden collections” of field books at the Smithsonian. The early years of The Field Book Project was composed of two phases. During the initial phase, FBP 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 which will be opened to other institutions to add catalog records for field books in their repositories.

Field books at the Smithsonian Institution are maintained by multiple departments or units: NMNH science Departments, Smithsonian Institution Libraries (SIL), the National Zoo, Smithsonian Tropical Research Institute (STRI) and SIA. Some departments, like the Department of Botany, have previously created inventory lists with basic information to help staff members and patrons find the field books they needed. In other Natural History departments, SIA created basic finding aids in the 70s and 80s and did some light processing of their collections, 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 method that combines descriptive practices from all three fields. Our project has drawn from metadata and encoding schemas and content standards from all three 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 rely upon other documents within the collection to provide context and tell a full story.

Libraries have long been innovators for item level metadata, which has traditionally emphasized access points like 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. These allow researchers to more easily pinpoint desired volumes. Prolific scientists may create over one hundred 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 also catalog at the item level following library descriptive practices.

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 expedition *United States Exploring Expedition* might consist of field books created by various individuals that participated in that expedition. Less frequently, our collections are assembled by the organizations as a creator. However they are grouped, collections are determined based on 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, in accordance with archival practice.

QUESTION 2—Digitization: Why should we digitize? How can digital versions of these materials enhance access to the material and 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 also recognized the need for remedial conservation so that many field books could survive the necessary handling required of this effort. Conservation tasks in the workflow needed to be a concurrent feature 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. This 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 digitization of text. For 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 best inform the process of searching for field book content. The catalog records we created employed 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 the 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 the vision of 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 their army of volunteers to perform transcription services. Our project competes with other SI projects for the attention of transcribers, and progress is slow. Other services must be considered if transcription of field books is going to become a critical and productive resource for biodiversity research.

QUESTION 3—Why is integrated access to data and images across scientific disciplines, institutions and type of resource (e.g. via aggregators like idigBio and BHL) 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 while 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.

QUESTION 4—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 SI Archives. From the beginning, the Office of the Chief Information Officer (OCIO) was engaged to help oversee the technical development and manage resulting content.

A Digital Asset Management System (DAMS) receives, stores, and replicates digital content that is produced by the Field Book Project. All catalog records, page level images and transcriptions are available to the public on the Smithsonian Collections Search site which contains information from all 21 SI museums and research facilities. FBP content is now available in the Digital Public Library of America and the Biodiversity Heritage Library.