MERGING METADATA: BUILDING ON EXISTING STANDARDS TO CREATE A FIELD BOOK REGISTRY

by Carolyn Sheffield, Sonoe Nakasone, Ricc Ferrante,

Tammy Peters, Rusty Russell, Anne Van Camp

Abstract

The Field Book Project is a cross-disciplinary project to develop an online registry for field books and other primary source materials related to biodiversity research. Led by Rusty Russell and Anne Van Camp, this project is a joint initiative of the Smithsonian Institution Archives and the National Museum of Natural History. This paper presents the metadata structure established for building the Field Book Registry. The project team is committed to involving members of the library, archives, museum, and biodiversity communities in the development of the Field Book Registry. We invite your comments and discussion regarding the work presented here.

Background

Field books are the primary source records of collecting events conducted for biodiversity research. In addition to the information that is typically transcribed to a specimen label, field books may also contain a wealth of other information useful to biodiversity research. Detailed notes on observations at the time of collection can provide clues about interspecies relationships. Environmental data can be used to guide habitat reconstruction and responsible land management. A less clinical yet highly valuable and intriguing contribution are the personal and cultural insights that can be gleaned from prose journal entries found in some researchers' notes. The goal of the Field Book Project is to expose these frequently overlooked primary source materials through the creation of a collaborative online Registry.

Statement of the Problem

Field books, despite their incredible research value, remain an obscure and difficult to access resource within many natural history collections. At the time of publication, the authors can find no evidence of standards of practice for providing access to field book collections and current efforts can be described as disjointed at best. Within a given institution, field book collections may be managed by multiple departments with no centralized access point. The unique nature of these textual objects suggests archival custodianship as the most logical, yet field books are just as frequently managed in museum collections, science labs, and discipline-specific libraries. These various types of custodianship result in myriad descriptive practices with varying levels of detail. Field books managed by archives are typically grouped together by creator or expedition and described in collection-level finding aids. Discipline-specific libraries produce item-level inventory lists with information similar to what might be found in a typical library catalog record. However, as these are unique and non-circulating items, these item-level records are not necessarily made available through the libraries' public facing catalogs. Finally, researchers wishing to access field books managed as part of museum collections and science labs will likely find themselves relying on the institutional memory of multiple staff members.

Within a given institution, the lack of an aggregated resource and prevalence of inconsistent documentation practices makes discovery – for both onsite and remote researchers – especially challenging. Field book collections housed in any given repository may represent collecting events from all over the world. Add to this the fact that a given collection – a set of field books from the same collector or expedition – may be distributed across multiple repositories and the process of locating all relevant materials quickly becomes frustrating. For this reason, it is especially important to bring together dispersed collections into an aggregated resource so that all research on a select geographic area can be accessed through one online location.

The biodiversity research community has already made great strides in developing effective collaborative, consortial information resources that we can model this project after. Some notable examples include: the Biodiversity Heritage Library (BHL)[1]; Encyclopedia of Life (EOL)[2]; Global Biodiversity Information Facility (GBIF)[3]; European Distributed Institute of Taxonomy (EDIT)[4]; and uBio[5].

Our Approach

Due to the complexity and multiplicity of the access issues surrounding field notes, an effective solution requires a multi-pronged approach. There is a need to: balance description between providing enough context for researchers to make relevancy assessments and supporting an efficient workflow for creating the catalog records; aggregate records for geographically dispersed collections through one online search interface; digitize and index page-level content despite frequently difficult-to-read handwriting; and establish the crucial links between the collecting events described in field notes and the resulting specimens and published literature.

This paper focuses on the metadata solution for the Field Book Project. In future publications, we will elaborate on the other aspects of our system design and architecture and the related work being performed under our sister project, Connecting Content.[6]

---

[1] BHL is a consortium of 12 natural history libraries that digitize the legacy literature of biodiversity held in their collections and provide open access as a global "biodiversity commons." http://www.biodiversitylibrary.org/

[2] EOL brings together several of the world's leading natural history institutions, botanical gardens, and libraries. http://www.eol.org/content/page/institutional_partners

[3] GBIF's mission is to make the world's biodiversity data freely and universally available via the Internet. GBIF provides a global informatics infrastructure for biodiversity research and applications worldwide. http://www.gbif.org/informatics/infrastructure/

[4] European Distributed Institute of Taxonomy is a consortium of 29 natural history institutions, in Europe and beyond, that provides tools for accelerating global production of taxonomic knowledge. http://www.e-taxonomy.eu/

[5] uBio is an international initiative within the science library community to create a comprehensive, collaborative catalog of known names of all living and once-living organisms.

http://www.ubio.org/index.php?pagename=general

[6] Connecting Content is a multi-institutional research project, led by the California Academy of Sciences and funded by the Institute for Museum and Library Services, to explore the connections between field book content, specimen collections, and published literature.

Field Book Metadata

The Field Book Registry is envisioned as a collaborative resource that will include catalog records for field book collections held in various natural history museums, libraries, and archives. Therefore, one of the primary goals is to provide a framework for description that a range of institution types and sizes will find easy to implement. It is equally important that the resulting system reflects community-agreed-upon methods for accessing these materials. To achieve this, a number of librarians, archivists, and natural history professionals convened to identify what they would like to see in a metadata schema for field books.[7] The following are some of the key requirements that were identified for the development of the metadata structure:

1. Support both collection-level and item-level description
2. Capture page-level metadata to enable navigation for future digitization efforts
3. Include data elements related to natural history collections
4. Include data elements related to text-based materials
5. Enable the creation and maintenance of authority files
6. Be based on available standards whenever possible
7. Be freely available and easy to implement within both large and small institutions

The project team worked with the same stakeholders to identify an existing metadata standard for this effort. A review of natural history, library, and archives descriptive standards revealed that no single pre-existing metadata standard would sufficiently meet all of the above requirements. However, there were several available standards from these three domains that, in combination with each other, would clearly meet the identified needs. Four schema were chosen: the Natural Collections Description (NCD); the Metadata Encoding Transmission Standard (METS); Metadata Object Description Schema (MODS); and Encoded Archival Context for Corporate bodies, Person and Families (EAC-CPF). Each of these is available as XML and is free to use.

Our approach will be to merge these, thus creating a data structure capable of providing a smooth transition from collection-level, contextual description to item-level access and will eventually support page-level navigation under future digitization efforts.

Sections 4.1 through 4.4 presents an overview of each of the four selected schemas along with descriptions of the value each contributes to the Field Book Registry. Section 4.5 demonstrates how the four schemas will interact to form one unified data structure.

---

[7] Participants in these discussions included representatives from the Biodiversity Heritage Library, Biodiversity (BHL) and BHL-Europe, Botany Libraries of the Harvard University Herbaria, California Academy of Sciences, Ernst Mayr Library at Harvard University, LuEsther T. Mertz Library at The New York Botanical Garden, Missouri Botanical Garden, the Royal Museum for Central Africa, and the Smithsonian Institution Libraries, Smithsonian Institution Archives, and Smithsonian Institution's National Museum of Natural History.

Natural Collections Description (NCD)

NCD is a metadata schema for covering all types of natural history collections, including text-based materials such as archives and published literature. NCD was developed by the **Biodiversity Information Standards (TDWG)**, formerly known as the Taxonomic Databases Working Group. The Field Book Project will implement NCD v0.7. This version offers a simple XML-structure that will integrate smoothly with the other XML-based schemas adopted for item-level description and authority file creation within the Field Book Registry.

NCD was selected for its ability to provide a rich yet approachable collection-level data structure. Field books and journals frequently comprise the core documentation of all collecting events from a given researcher's career. As such, collection-level description helps to maintain the functional context in which each volume/item was created and establishes clear relationships to other items created within the same context. In addition, description at this level is an efficient way for institutions without the resources to perform item-level cataloging to begin to describe and provide access to their collections.

Being a schema developed for covering natural history collections of all types, NCD is also well-suited for the expansion of the Field Book Registry under the Connecting Content project. Its extent is capable of supporting the development of a rich network of relationships across collection types held by multiple repositories. This will be especially useful for enabling cross-searches of specimen and published literature collections.

Metadata Encoding Transmission Standard (METS)

METS[8] aggregates descriptive, administrative, and structural metadata for digital library objects. The standard is maintained in the Network Development and MARC Standards of the Library of Congress, and is being developed as an initiative of the Digital Library Federation.

METS was selected for its ability to support sequential page navigation for digitized books; an important extension for the Registry once digitization efforts are underway. Hyperlinks to METS records for the individual field books comprising a collection (e.g., same collector or same expedition) will be listed in an umbrella NCD record. This collection-level grouping of item-level records will help add valuable context for remote researchers trying to determine the significance and relevance of each item. For item-level description, the Registry will include MODS records within the descriptive wrapper of the METS framework. (See Figure 1).

Metadata Object Description Schema (MODS)

Developed and maintained by the Library of Congress, this schema consists of a subset of MARC[9]-compatible fields as XML-tags, making it possible to easily map from one schema to another. Additionally, MODS[10] tags are language-based rather than numeric codes found in MARC 21 (Guenther and McCallum, 2003). This provides a notable advantage for the multi-institutional Registry as some contributing institutions may not have individuals on staff trained in traditional library cataloging.

---

8 To view the schema and documentation, please visit http://www.loc.gov/standards/mets/mets-schemadocs.html
9 Machine Readable Cataloging is an encoding standard for bibliographic description used in libraries around the world. Current version is MARC 21: http://www.loc.gov/marc/
10 To view the schema and documentation, please visit http://www.loc.gov/standards/mods/mods-schemas.html

A primary objective of this project is to build on existing standards to bridge the collection- and item-level metadata gap. Due to the nature of the materials, and their relationships to a vast number of other items in natural history collections, effective and efficient access benefits from item-level description. MODS captures item-level descriptions and will be used within the METS descriptive metadata wrapper. As MODS is largely based on MARC 21, crosswalks are already available for ingesting data that may already have been populated in a MARC catalog. In addition, since all of the schemas are XML-based, developing metadata crosswalks between MARC, Dublin Core and any number of other standards to populate the descriptive wrapper. Zarazaga-Soria, et al. (2003) describe a process for effectively establishing and maintaining crosswalks using XSLT and a metadata crosswalk broker.

Encoded Archival Context for Corporate bodies, Person and Families (EAC-CPF)

The EAC-CPF schema enriches the context of historically significant collections by capturing information on the entities involved in the creation, use, and maintenance of the materials. EAC-CPF is maintained by the Society of American Archivists in partnership with the Berlin State Library[11]

The Field Book Registry will use EAC-CPF to ensure consistent and controlled entry of names related to the creation and maintenance of the collections. This module will provide historical and bibliographic context and help reduce ambiguity for person and corporate names. During the creation of these authority records, we will consult resources such as the Virtual International Authority File (VIAF)[12], author list from the International Plant Names Index (IPNI), botanist database of Harvard University Herbarium, and other sources in which names are standardized.[13]

The project team also hopes to benefit from a similar effort underway in the Smithsonian Institution Archives to create Expedition Histories, historical context records about expeditions. Although created in MARC, the underlying concepts included in these Expedition Histories are similar to an EAC record and could be mapped to an EAC instance.

Creating One Data Structure

Our approach is to bring together the key data elements and relationships from each schema to form one unified metadata solution. Figure 1 shows a select set of elements from each schema to illustrate the points at which the four connect. In the diagram, each box represents one of the four schemas. The elements listed inside the boxes represent only a small subset of those retained for use in the Field Book Registry. These elements were selected for inclusion in the diagram because of the role they play in connecting the four schemas to one another, as indicated by the lines connecting elements and schemas. Since we are working directly in a hierarchical XML structure, it is important to emphasize that these elements should not be confused with columns in relational database tables.

---

[11] For more information, please visit http://eac.staatsbibliothek-berlin.de/ (http://eac.staatsbibliothek-berlin.de/)

[12] From their home page: "VIAF, implemented and hosted by OCLC, is joint project of several national libraries plus selected regional and trans-national library agencies." http://viaf.org/

13 Additional resources include Taxonomic Literature 2 (TL2), Smithsonian Annual Reports, Web of Science, Encyclopedia of Life, and membership lists of scientific societies. We anticipate that this list will grow as partners begin contributing records to the Registry.

Relationships are formed between elements of the different schema and, in some instances, between elements and an entire schema. Starting with the collection record, NCD contains the elements, *Collector* and *Owner*, which are populated with links to the *nameEntry* element of the EAC records. This ensures normalized spelling and enables entity disambiguation. This is especially important if two collectors have the same name, as an EAC record can be used to distinguish on the basis of life dates and other biographical data. For persons, the name element in an item-level MODS record will also point to the EAC record. In the example shown, the person in the MODS record also has the role of creator and therefore points to the same EAC record as in the NCD example. This is important because a collection, particularly if grouped by expedition, may include field books created by multiple persons, only one of which may be relevant at the item-level. On the other hand, institutions in the role of owner or custodian are not linked from the MODS record to EAC. Rather the *physicalLocation* element (not shown) in the MODS record is populated with an URL pointing to the institution's website.
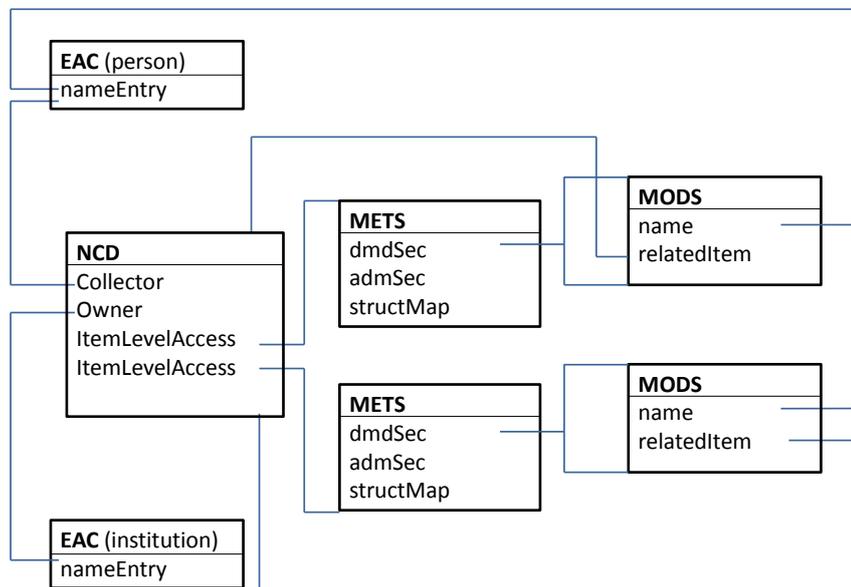

Figure 1: Relationships between the four selected schemas for the Field Book Registry.

The merged schema also supports the reciprocal relationship between collections and items. The element *ItemLevelAccess* points from the collection-level NCD to the item-level METS record. The metadata wrapper, *dmdSec*, contains a descriptive metadata record within the METS record. For the Field Book Registry, this will be a MODS record. MODS includes the element *relatedItem* which will point back to the NCD record. By capturing the relationships at both the collection and item levels, the Registry provides end users with the ability to navigate to broader or narrower results sets and back again.

Assessment of Metadata Structure

The authors acknowledge the complexity of merging four distinct metadata schemas into one resource. To support our project goal of creating an approachable and easy-to-implement system for a range of institutions, we have assessed each of the four schemas and identified areas for simplification.

Methodology

Project staff worked with internal and external partners to conduct a number of reviews of each of the four schemas. The first review involved a manual review of data definitions for all elements in each of the four schemas. As part of this effort, professionals involved in the creation and maintenance of the four schemas were contacted for additional information on data definitions and intended use. Once a preliminary set of changes had been implemented, test records were created. These test records were assessed for impact on workflow and additional areas were identified for modification. The records were then shared with internal and external partners for further review and suggestions for streamlining.

It is important to note that elements determined to be irrelevant for our project will not be deleted from the XML schema in the system's backend. Rather this information will be excluded from cataloging templates, thus ensuring flexibility and allowing other institutions to reinstitute any of the elements for their own needs.

Results

In reviewing the elements, three main criteria for determining inclusion versus elimination emerged: (1) occurrence; (2) granularity; and (3) relevance to material type. We also identified certain elements that would require more specific guidelines to ensure consistent data entry practices in a multi-institutional resource.

Occurrence

Although none of these schemas alone is as robust as MARC, together they present an overwhelming number of fields and choices. For this reason it is necessary to reduce the occurrence of elements that capture essentially the same information albeit in different places. The elimination of redundant elements is best illustrated by examining the significant overlap between NCD 0.7 and EAC.

NCD 0.7 is comprised of three main sections: *Collections*, *Persons*, and *Institutions*. The latter two serve essentially a similar purpose and collect largely the same data as an EAC record for a person or corporation. Not only is the overall purpose essentially the same, but the data collected is largely the same. One notable difference is the use of a *vcard* namespace for NCD. As the Registry will point users directly to the custodian institution's website, we did not find this level of granularity necessary. Overall, we found EAC to more closely align with what we were trying to achieve in creating contextual information about those creating, using or maintaining field books. Therefore we will not be using the *Persons* and *Institutions* sections but rather providing links to EAC records for organizations and people in the NCD fields *Owner*, *Collector* and *AssociatedPerson*, respectively. Some redundancy has been maintained in order to facilitate linking between the four schemas.

Granularity

Even with the reduction of redundant elements, there are still an immense number of elements to contend with. To move closer to an efficient workflow, we eliminated or restricted use of certain fields that would present an unnecessary or unrealistic burden on catalogers. For example, EAC provides the following fields in which descriptive information for an entity is recorded: *structureOrGenealogy*, *places* (of existence or activity), *localDescription*, *generalContext*, *function*, *mandate*, *occupation*, *existDates*, *legalStatus*, and *bioHist*. Rather than using all of these fields, the *bioHist* field will be used for a free text general description about the entity. *ExistDates*, *occupation*, and *legalStatus* (for corporations) will be retained to provide a brief overview of the entity.

Relevance to field books as a material type

Due to the archival nature of field books (i.e., unpublished, primary source documents) certain elements from the library-based MODS schema and the natural history-based NCD schema were found to be irrelevant. For example, *originInfo* in MODS largely supports information related to the publication of a text. With the exception of *dateCreated*, we have eliminated all subfields under *originInfo*.

Similarly, while field books often describe specimens, not all specimen-related metadata will be relevant. In NCD, *SpecimenPreservationMethod* and *ConservationStatus* both specifically refer to the actual specimens and will not be used in the Field Book Registry.

Refining Guidelines

From the remaining element set that would form the Field Book Registry, some elements required refinement to ensure consistency across multiple institutions. In addition to using normative examples provided by the schema and content guides such as AACR2 (2005) and DACS (2004), we strove to remain as consistent as possible with practices already in place at partner institutions. Working closely with library staff at the California Academy of Sciences (CAS), we reviewed California Digital Library (CDL) guidelines and sample CAS records to guide decisions on how to format entries in date, name, and language fields. Additionally, CAS practices informed our decision to use AAT as a primary source for genre terms.

Conclusions / Future work

A primary goal of the Field Book Project is to ensure that participation in the Registry is a realistic prospect for a range of institutions. For this reason, we plan to conduct workflow studies with partner institutions to evaluate the impact of the metadata schema on their cataloging workflow. We will work closely to assess the data structure for learnability, intuitiveness, and workflow efficiency as they create their own use case records.

We also look forward to establishing rich networks between field books and other collection types under the Connecting Content project. Several pilot projects will launch in 2011 to explore these ideas and we anticipate exciting and informative results.

References

Guenther, Rebecca, McCallum, Sally. New metadata standards for digital resources: MODS and METS. *Bulletin of the American Society for Information Science & Technology*, Dec 2002/Jan 2003. Last accessed January 28, 2011 at: http://findarticles.com/p/articles/mi_qa3991/is_200212/ai_n9150534/

Zarazaga, F.J., Torres, M.P., Nogueras-Iso, J., Lacasta, J., Cantan, O. 2003. Integrating geographic and non-geographic data search services using metadata crosswalks. *9th EC GI & GIS Workshop, ESDI Serving the User, A Coruña, Spain.* Last accessed January 28, 2011 at http://www.ec-gis.org/Workshops/9ec-gis/papers/services_torres.pdf

Natural Collections Description. Developed by TDWG Interest Group. Overview: http://www.tdwg.org/activities/ncd/ and v0.7 Schema: http://rs.tdwg.org/ncd/0.70/ncd.xsd, last accessed January 28, 2011.

Encoded Archival Context-Corporate Bodies, Persons and Families. Maintained by the Society of American Archivists in partnership with the Berlin State Library. Last accessed January 28, 2011 at http://eac.staatsbibliothek-berlin.de/

Metadata Object Description Schema. Developed and Maintained by the Library of Congress. Last accessed January 28, 2011 at. http://www.loc.gov/standards/mods/

Metadata Encoding and Transmission Standard. Maintained in the Network Development and MARC Standards Office of the Library of Congress, and developed as an initiative of the Digital Library Federation. Last accessed January 28, 2011 at http://www.loc.gov/standards/mets/

Describing Archives: A Content Standard. 2004. Chicago: Society of American Archivists.

Anglo-American Cataloguing Rules, Second Edition, 2002 revision, 2005 update. *Published Jointly by the American Library Association (ALA), the Canadian Library Association (CLA), and the Chartered Institute of Library and Information Professionals (CILIP).*

CDL Guidelines for Digital Objects (CDL GDO). Version 2.0 Maintained by the California Digital Library Reviewed and Updated Semi-Annually. Last accessed January 28, 2011 at http://www.cdlib.org/services/dsc/contribute/docs/GDO.pdf