

Increasing Online Discoverability of a Mixed-Format Collection

The Collection

In 2004, Smithsonian Libraries acquired the mixed-format Russell E. Train Africana Collection for its special collections division. This collection contained items that had broad public appeal and significant historical value. The collection's diversity of materials has been a source of excitement and challenge since Smithsonian Libraries acquired it in 2004. Judge Russell E. Train created the collection around his decades-long fascination with the history of exploration and wildlife in Africa.

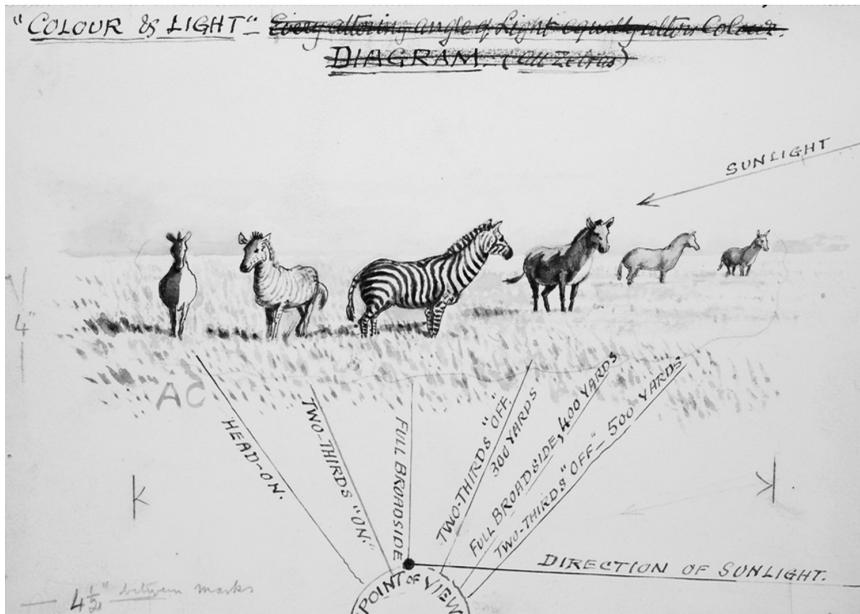


Image 1. Sketch by Abel Chapman, 1851–1929. Part of "Sudan sketchbooks 1913–1919," Art022. Part of the Russell E. Train Africana Collection at the Joseph F. Cullman 3rd Library of Natural History, Smithsonian Libraries. SIL28-298-01.

Train acquired materials from legendary figures like Theodore Roosevelt during his African Expedition in 1909–1910, as well as explorers David Livingstone and Henry Morton Stanley. The collection includes documents, photographs, and personal items from more recent figures, like Ernest Hemingway, when he was on safari during the 1950s.

Questions arose concerning how to best describe the content and make that description discoverable to as wide an audience as possible, while still adhering to library and archival standards. Books in the collection were simple to address, as they could be cataloged following the existing rare books workflow. However, the nontraditional library content, including manuscripts, photographs, original artwork, posters, maps, and ephemera, presented additional questions.

The collection creator, Russell E. Train, had supplied a detailed inventory of these materials in a text document along with the collection. The Smithsonian Libraries wanted to make these descriptions in the inventory widely available and to supplement, expand, and—if necessary— reformat them to increase their utility for researchers. In the spring of 2015, Smithsonian Libraries decided to pursue a short-term project to make the description available, while taking into consideration interoperability with the publishing platforms, both internal and external, to which the library had access. Through this process, staff dealt with questions relating to the needs unique to digital content, challenges of data reuse, and descriptive standards.

The libraries initially looked at two options: create records for each item or a finding aid for the collection. Records for each item would take advantage of the wide range of online catalogs to which Smithsonian Libraries contributes, like Biodiversity Heritage Library and Digital Public Library of America. Smithsonian's Office of the Chief Information Officer started providing support to institutional departments and divisions that wish to use Encoded Archival Description (EAD), the archival profession's standard descriptive schema. This support made a finding aid more plausible and more appealing. An online finding aid would enable users to search the Train Collection alongside other Smithsonian archival collections, as well as show collection- and item-level relationships more clearly.

Over the course of the project, Smithsonian Libraries created 500 item-level MARC records and an online EAD finding aid. For many libraries, this would be potentially problematic. Item-level records offer more information than a box list in a finding aid, so why risk confusing users by essentially offering duplicate records in two schemas (item-level MARC records and an EAD finding aid)?

“Crosswalking” EAD and MARC

Special collections and archives have recognized the increasing need to reach audiences beyond the traditional “silos” of library and archive platforms. Since the early 2000s, there has been work conducted, like Katherine M. Wisser and Jennifer

O'Brien Roper's project in 2003¹ at North Carolina State University Libraries, to crosswalk from EAD to MARC. Since then, there have been articles² discussing in greater depth how to bridge multiple schemas, as well as repurposing and reusing existing description. This idea of reusing description or metadata has now become part of the data curation conversation. The Research Libraries Roundtable panel on Data Management and Curation in 21st Century Archives³ at the 2015 annual meeting of the Society of American Archivists in Cleveland, Ohio, covered this very topic. However, these discussions usually look at how to repurpose EAD into a MARC collection-level record, not the item level, or they avoid the challenges of cataloging nontraditional library materials in MARC by using other schema.

Institutions have taken advantage of methods that harvest from repositories compliant with the Open Archive Initiative (OAI) and that use Dublin Core, like the project at Brigham Young University described by Corey Nimer in 2011,⁴ which allow them to automatically contribute records to organizations like WorldCat. Smithsonian Libraries workflows and online relationships made some of these options unworkable. Smithsonian Libraries has a well-developed digitization workflow for MARC records; the libraries only use OAI-compliant systems for their repository. Records not in MARC could be contributed to Smithsonian's Collections Search Center, which accepts any type of record, but not to Smithsonian Libraries' primary record destination, SIRIS, which is a MARC-based system.

Complications and Possibilities

When the collection was acquired, the Special Collections department considered cataloging the materials using the rare books workflow, but the unique qualities of the archival content (unpublished, little detail provided, difficult handwriting) made cataloging too time-consuming. They looked at hiring an archives contractor to reorganize the archival materials. However, Special Collections decided against this. Traditionally, finding aids and archival description focus on the collection and series level, providing context for how the materials relate to each other. However,

1. Katherine M. Wisser, Jennifer O'Brien Roper, (2003). "Maximizing Metadata," *Association for Library Collections and Technical Services Journal* 47, no. 2 (2003), doi:10.5860/Irts.47n2.71.

2. Heidi Frank, "Augmenting the Cataloger's Bag of Tricks: Using MarcEdit, Python, and PyMARC for Batch-Processing MARC Records Generated From the Archivists' Toolkit," *Code{4}Lib Journal*, issue 20 (2013), available online at: <http://journal.code4lib.org/articles/8336> [accessed <<need access date>>]; Teresa M. Keenan, "Why Purchase When You Can Repurpose? Using Crosswalks to Enhance User Access," *Code{4}Lib Journal*, issue 11 (2010), available online at <http://journal.code4lib.org/articles/3604> [accessed <<need access date>>].

3. Ixchel M. Faniel, Wendy Hagenmaier, Sammie L. Morris, "Data Management and Curation in 21st Century Archives," Research Libraries Roundtable panel at the Annual meeting of the Society of American Archivists in Cleveland, Ohio, August 2015.

4. Corey Nimer, "Applying Inheritance: Single-level Displays and Repurposeable Metadata," Society of American Archivists 2010 Research Forum (2011), available online at <http://www2.archivists.org/sites/all/files/CNFinal.pdf> [accessed <<need access date>>].

this particular collection is artificial in structure, since the whole did not originate organically from a specific function; thus, the materials do not necessarily gain additional meaning from the other items. Instead, the libraries hired a contractor to migrate the data from the inventory to an Access database. The Special Collections department revisited the description challenge in 2015, when Smithsonian Institution's Office of the Chief Information Officer began to offer support and training in the use of EAD for the purposes of Smithsonian Online Virtual Archive.

By this time there were significant developments in the online environment. These complicated the discussion about how to best make these materials discoverable. Several online platforms such as Internet Archive, Digital Public Library of America, and Biodiversity Heritage Library to which Smithsonian Libraries now contribute did not exist or were in their infancy. Online catalogs that aggregate EAD finding aids were still in development. When the collection was acquired, reference requests were usually directed to Special Collections staff, so the detailed inventory from the collection creator was often sufficient to answer questions.

As digitization increases and online platforms multiply, researchers may end up never interacting with special collections staff for access. The inventory met basic needs for searching the physical materials, but the special collections staff is no longer dealing with just the physical object. Some of the collection materials have been digitized and are available online; more will be digitized. Additionally, the digital object is not constrained in the same way as the physical object. The physical materials reside in a special collections library, but the digitized versions would make sense in online platforms containing materials from libraries, archives, or museums. Smithsonian Libraries wanted to create records to address the needs of both the physical and the digital materials.

The materials in question could be described in several different descriptive schema based on their physical location (in a library) and the nature of their content (original materials). The libraries decided that the greatest priority was to make the description available in as many places as possible. User confusion could be kept to a minimum as long as staff used best practices for record creation whenever possible. The project coordinator would work with staff from Smithsonian Online Virtual Archive and the cataloging staff to determine what descriptive standards would be used. The resulting item records would meet two needs: items could be found online and provide ready-made metadata for digitization. The finding aid would display the hierarchical relationship between materials and the collection context.

So how did Smithsonian Libraries decide on MARC and EAD?

EAD is clearly the standard for archival collections, but there are numerous standards that could be used for item-level records, especially if the records are to be used as metadata. Given the variety of formats, Dublin Core, a schema designed to be “broad and generic... for describing a wide range of resources”⁵ would surely be the most logical choice. These questions springboard into additional queries complicated by differences in archive and library description methods.

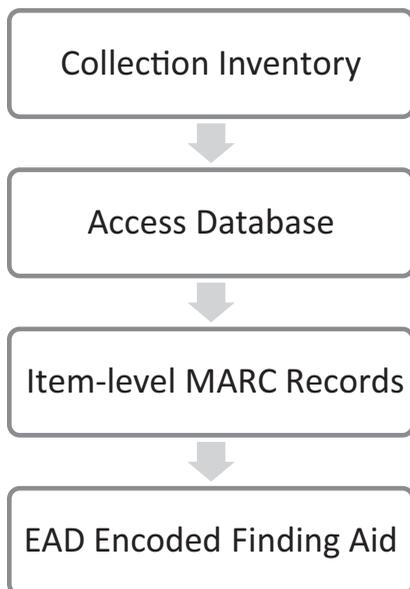
Smithsonian Libraries looked at the current organization and determined that the existing structure in the Access database was sufficient. Following this line of reasoning, project staff determined to make decisions that would make the best use of existing description and structure. The existing structure fit better with MARC than Dublin Core, since the Special Collections cataloger at one point had augmented some records with subject headings and other MARC-based description. It also meant records would be in the department’s preferred schema (MARC). Existing item description was robust. Staff thought about splitting apart collection “items” that might normally be cataloged separately. Project staff determined the priority was to make the most of existing description to make materials discoverable and would not use limited staff time to reorganize the collection to meet a standard cataloging definition of “item.”

The Project

The Smithsonian Libraries started the three-month project with a team that included members of the cataloging department, digital services, special collections, and a project coordinator who worked on another Smithsonian project between the Institution Archives and the Libraries. This team investigated how to best use the existing description in the Access database created from the collection inventory. This involved looking at what standards to use, where to contribute, and how to make the records sustainable; that is, how to take into account that some online publishing platforms would require manual updating of the records if they were changed in the future. This last concern was particularly important because the Smithsonian Libraries are very active in contributing to online consortia. If content was contributed to one location, it would be automatically uploaded to additional sites over which the Libraries had little editing control.

Due to the range of possible solutions, the project coordinator also worked with staff at OCLC, the Smithsonian Libraries DAMS coordinator, and the Office of Chief Information Officer, to name just a few. To inform these decisions, the project team compared the existing description to online catalog requirements; what standards they wanted to meet (library and archival) and how much editing would

5. “Dublin Core Metadata Element Set, Version 1.1”. Dublincore.org. Retrieved 5 April 2013.



be needed; what online locations would be both desirable in terms of reaching new audiences and sustainable in requiring minimal future editing; and how to fit this work into existing workflows so that records would be used by staff and easily maintained after the project was finished.

Image 4. Record Creation process. Inventory was supplied by the Collection creator. Smithsonian Libraries hired a contractor to convert the inventory into an Access database. The short-term project used the database description to create item-level MARC records and the box list for the Finding Aid.

Accessing Online Locations in Light of the Digital Item

The first step in this cataloging project was to select ideal platforms where the new collection records would be made available. Staff determined that there were three ideal destinations for records: the institutionwide online catalog Collections Search Center (CSC); the Smithsonian Libraries online catalog Smithsonian Institution Research Information System (SIRIS); and Smithsonian’s Online Virtual Archive (SOVA). To determine the costs and benefits of each, the project team assessed current workflows as well as what changes and additional time would be necessary to bring a set of records to each platform. This process proved illuminating in terms of understanding that many staff would be needed to contribute and maintain records in each location unless the work was carefully structured to use as much machine effort and as little staff effort as possible.

The project team determined that the following principles would guide the creation of new records and workflows for online publication:

- Creating “new” records would be defined as transforming existing item-level descriptions into currently used library and archive standards (such as RDA, EAD, and MARC)
- When creating records, existing item descriptions would be used without significant edits to save time
- Current department workflows should be used for creating and uploading records when possible—records would therefore reside in expected locations, and future edits, corrections, or additions could follow current workflows

- Assign structures and data fields to the records so that they are searchable in a way that will make sense to researchers and to the library staff that assist them
- Ensure that all the records for the entire collection are available and searchable in one place, and/or are collocated within each online platform
- If possible, the resulting description would be used to make archival portions of the collection available alongside other archive collections at Smithsonian through contribution to Smithsonian's online virtual archive (SOVA)

Of the three ideal publication platforms identified, the easiest to use was the Smithsonian Collections Search Center (CSC). The CSC is format-neutral, such that the records could be converted quickly into Dublin Core and all of the data would immediately be surfaced through CSC to the public. However, the CSC is an endpoint, a final resting place for catalog records that do not contain digital content, so contributing first to CSC would mean that the records would be in a fraction of possible locations and would have to be managed outside the cataloging workflow.

If, on the other hand, records were created in MARC and entered first into the library's online catalog, there would be significant benefits. Records would be entered once and published many times—they would automatically populate anywhere Smithsonian Libraries already contributes MARC records (including CSC and OCLC). Records for the entire collection would therefore exist in more than one location, vastly increasing discoverability. At the completion of the project, records could be managed and searched in the same manner as any other records. Records would be aligned with other internal workflows such as those for digitization, which uses the item's MARC record from the library's online catalog and meant that any future digitization requests could be fulfilled with a minimum of effort using existing workflows.

Reusing the Description

The project team felt that, for this collection, item-level records in MARC could be created using the content provided by the collector. The descriptions provided by the collector were very detailed and already loaded into a database, which would make reuse and transformation much faster. All the collection holdings could then be searchable at the "item" level together in the online catalog—an exceedingly appealing possibility for the research user community.

However, the database content did pose some challenges. The meaning of "item" varied. In terms of library cataloging, each item ideally receives its own record; however, "items" in this collection were more like "folders" in archival terms. The "items" described in the database were sometimes small groupings of materials that were purchased or combined by Train. Content did not appear to meet one

descriptive standard, such as library standard AACR2 or archives standard DACS. Description varied. Some items included abstracts and biographical notes, while other materials had no identified creator.

To use the database, library staff decided to use the existing level of description; “item level” would essentially mean “folder level.” Whatever grouping level Train maintained would be the level of description. This would free up staff time to focus on the greater goal of the project—getting these records to the public via an online discovery platform. This also worked well toward the goal of creating a full finding aid. Since the project would be using the folder level as item level, item-level records could be transformed into a box list for the finding aid.

Project staff then focused on accessing the content and quality of the existing description. They determined which parts of the database content correlated with MARC fields. They also looked at records already created by the special collections cataloger for archival items in the collection, and mapped out how the database descriptions would be used to create minimal MARC records.

Descriptive Standards

Two ideas guided many of the decisions made in terms of record editing. First, as these records are unique, there is little chance that they will be copied. For this reason, staff focused on only editing or adding content if it affected the searchability and accessibility of the record. Second, these would be MARC records, but they would not be entirely DACS, AACR2, or RDA compliant. Changes focused on clarifying description by removing abbreviations and terms specific to rare books. This was also done in the spirit of RDA, which often focuses on spelling out terms and transcribing what is on the item. Text in title and summary was largely left unchanged. Special care was given to standardize access points like creator, related names, and subject headings to comply with the International Standard Bibliographic Description (ISBD). Subject headings taken from the collection record were added to all item-level records. Since there are so many different formats, staff standardized format description with the Getty Research Institute’s Art & Architecture Thesaurus. Changes were made to maximize searchability alongside other records, using carefully chosen professional standards.

The Results

The project was able to create 500 item-level MARC records and a finding aid in three months by working with staff who had expertise in library and archive description, with a goal of adapting existing description. These records also go well beyond minimum descriptive standards. The MARC records have been well received by the special collections department and the finding aid has received

compliments from across the units. These records have already enabled staff to make digital content more widely available. These include not only fully digitized versions of materials, but also images originally taken for conservation purposes that are now available on the finding aid in Smithsonian Online Virtual Archive.

Burroughs Wellcome and Company medicine chest used by Robert Henry Nelson, possibly during his time with the Emin Pasha Relief Expedition.

Add to my list

Author: Burroughs Wellcome and Company, manufacturer.
 Title: Burroughs Wellcome and Company medicine chest used by Robert Henry Nelson, possibly during his time with the Emin Pasha Relief Expedition.
 Publisher: [188-?].
 Description: 1 item ; 2.75 x 8 x 1 in.
 Notes: Supplied title.
 NelA003 is the accession number in the Russell E. Train inventory list of the Joseph F. Cullman 3rd Library of Natural History. Forms part of: Russell E. Train Africana Collection (Smithsonian Libraries)
 Summary: Burroughs Wellcome and Company medicine chest, with ten vials still containing remnants of various drugs.
 Location: DSI-NH Joseph F. Cullman 3rd Library of Natural History National Museum of Natural History, Room CE-G15 Constitution Avenue and 10th Street, NW, Washington, DC, USA 20013-7012
<http://library.si.edu/libraries/cullman>
 Subject: Emin Pasha Relief Expedition (1887-1889)
 Medical instruments and apparatus -- Specimens.
 Genre: Artifacts (object genre).
 Added Author: Nelson, Robert Henry, 1853-1892, former owner.
 Russell E. Train Africana Collection (Smithsonian Libraries)
 Catalog Source No.: (OCoLC)ocn910941768

Image 5. Example of MARC item-level record available through Smithsonian Libraries Online Catalog at <http://siris-libraries.si.edu/>.

Records also meet the other characteristics deemed necessary by the project team. Because staff used the prevailing cataloging standard for which existing workflows were designed, they benefited from Smithsonian Libraries' relationships with on-line platforms and the increasing institutional support of EAD. Because the records were created in MARC, the Train archival collection item-level records are now part of the regular cataloging department workflow for any future edits and contributions. Staff will be able to quickly meet any digitization requests for the unique materials, because MARC records already exist for needed metadata. The entire process has been fully documented and can be edited to convert and contribute other nontraditional materials in library holdings.

Though records were not cataloged to the standards of the rare books workflow, they do provide a strong alternative to standard cataloging for special collections. The archival community has developed the approach of "more product, less process," for processing collections. Who is to say that the expedited approach can't be applied to other library activities, such as more granular (series- and item-level) record creation? The goal is to make materials discoverable through the judicious use of staff time and expertise. Many libraries already have some type of description that accompanies new collections for inventory purposes. These documents can provide a framework for item-level record creation for both internal and external (online) use. Once the collection has basic item records in a familiar format like MARC, it can become even easier for catalogers to strategize which records

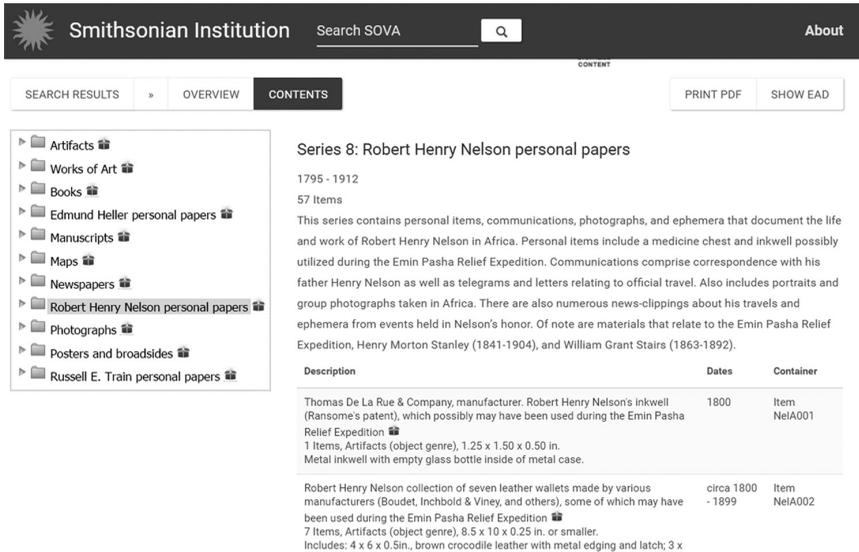


Image 6. View of Series 8 from collection finding aid available on Smithsonian Online Virtual Archive at <http://sova.si.edu/>.

are deserving of additional descriptive work. Abbreviated item records also gain powerful meaning when they enable digitization and are published online alongside digitized materials. When researchers can see the resources for themselves and verify the content, the possible downsides of abbreviated item description may prove negligible.

Lessons Learned

With so many possibilities, it is easy to forget that the technology and systems are tools. They should not overshadow the project goals. Our goals were to make the most of existing description in a way that can be easily managed in the long term, as well as take full advantage of existing workflows and institutional relationships for record distribution. We selected schema and descriptive standards for their longevity and adaptability with Smithsonian Libraries' current systems.

MARC may seem problematic because of its wealth of standards and requirements. However, these can be treated as guidelines and best practices. It is important to determine what is "good enough"; description that is sufficient to ensure staff can find and identify materials and that researchers have enough description to determine if materials may be pertinent to their research. By focusing on what makes the items unique, and their access points, we were able to make the most of staff time when editing.

The catalogers on our project team took on the important role of subject matter

experts, able to advise what was necessary in the records, what description had to be standardized, and what could be left alone. This allowed a cataloger whose primary background is not in MARC to complete much of the cataloging work and optimize the Special Collections Cataloger's time on the project.

If a library decides to pursue data reuse for records in multiple schemas, the libraries must be aware of which description is primary. Staff will not have time to update both if there are changes in the future. Since we are, by definition, a library, and our staff and researchers default to our MARC records, we created the primary description in MARC. Since we do not anticipate substantial content changes, we can embrace the opportunity to make description available in EAD, which only creates more connections and access points to the materials. Archives have embraced EAD as a way to finally make available archival collections in a manner that is searchable online and represents an item's context. Libraries with nontraditional materials and archival collections like the Train Africana collection can significantly benefit by participating in EAD's growth. Repurposing MARC records makes that possible in the face of staff and budgetary constraints.

The process was helpful in other ways. Library staff is now more familiar with archival description through EAD with the guidance of Smithsonian Online Virtual Archive (SOVA) staff. Cataloging staff were able to edit and fill out descriptions more easily when the information was put in a format they are used to (like MARC). This process of targeting the type of editing done to record description is now informing other Special Collections cataloging projects in the works.