Scurlock Photographs Cataloging Analysis

Adam Mathes
Administration and Use of Archival Materials - LIS 581A
Graduate School of Library and Information Science
University of Illinois Urbana-Champaign

December 2004

1 Executive Summary

The Addison Scurlock Studio Photographs and Records collection will be scanned and cataloged. The problem to analyze and put forth a recommendation on is whether this collection will be cataloged using traditional MARC records, or the more recently developed Encoded Archival Description (EAD) format.

This is a problem of archival and bibliographic description that encompasses technical differences between the standards, as well as understanding the larger problems and processes of archival cataloging, records creation and the design of finding aids. Metrics that need to be examined include ability to meet project goals, interoperability with other systems and institutions, future prospects of both formats, as well as resources necessary for each practice.

Although the specific advantages and disadvantages of both formats will be discussed, overall the question itself establishes a false dichotomy, and the recommendation is not to fully wed the project to either format, but catalog the photographs in such a way that both formats can be used effectively in the right contexts.

2 Institutional Context

No cataloging project takes place in isolation, and any analysis of the problem would be incomplete without understanding the collection and what the purpose of the project is. The collection in question is the Addison Scurlock Studio Photographs and Records, at the National Museum of American History. These are photographs, some already deteriorating, primarily reflecting city life in Washington DC, including life at the turn of the century, and of many notable African-Americans.

These records will need to be used for web accessible, searchable finding aids, printed inventories and finding aids, as well as integrated into the existing institution-wide catalog, the Smithsonian Institution Research Information

System (SIRIS). These can be grouped into two distinct but similar groups: records for catalogs - data - and inventories and finding aids - documents. This distinction is important as the information captured will need to be made useful as data for computers, as well as structured into reasonable documents for humans.

3 Models, Formats, Standards

In analyzing the problem, some basic terminology and concepts are essential background. When discussing the merits of metadata formats - that is, the way in which we represent data about documents, data, collections, and other artifacts - there are different levels at which decisions are made. There is a distinction between a **conceptual model**, an **encoding format** and a **content standard**.

- A **conceptual model** is a high level, abstract idea of what to represent with the metadata or cataloging records. For example, in this project, expressing that items are organized in series, which are part of collections, and that each item may have one or more named creators associated with it.
- An **encoding format** is a formal structure, or container, for expressing and representing that conceptual model. These are usually very technical specifications, that say in a formal, machine readable way things like "within the element called collection, there can be zero or more series elements."
- A **content standard** is a set of rules or guidelines that clearly specifies what goes in the containers specified by that encoding format. They specify things like how to represent an individual or corporate entity's name.

MARC and EAD are primarily encoding formats. They specify a format for encoding records in a rigorous, machine readable way. There is a conceptual model behind both of these formats, and while similar, they differ in important ways and have different origins. MARC is very closely associated with the content standard Anglo-American Cataloging Rules (AACR2.) EAD's relationship to content standards is more nebulous, and will be discussed as well. Additionally, these encoding formats are generally associated with particular processes and software, but those relationships need not always be true.

4 Encoding Formats

4.1 MARC

The Machine Readable Catalog (MARC) format is a thirty year old database structure originally intended to facilitate the printing and cataloging of bib-

liographic catalog cards in libraries. (Hensen 2001) This origin is important: MARC was not originally devised for archives, but for libraries, and was developed by the library of congress. Specifically, its origin is tied to the card catalog, and now to the Online Public Access Catalog (OPAC) systems that replaced them. This is a fundamental difference since in archives the primary access tool is often not an institutionally encompassing item-level catalog, but finding aids for specific collections.

MARC does have a history of use in archival institutions, beginning with MARC-AMC in the 1970's. (Hensen 2001) MARC is an encoding format, but is very closely associated with the AACR2 content standard. The archival community developed a complementary content standard, Archives, Personal Papers, and Manuscripts (APPM), that elaborated on AACR2 for archival specific contexts and artifacts.

Strengths of MARC include:

- Established, proven format - it's over 30 years old and still in use.
- Proven interoperability with libraries and other institutions - it was designed as an interchange communications format.
- Robust software systems exists to support creating and managing MARC records.
- Connection to well-defined content standards - AACR2 and APPM.

Some weaknesses of MARC include:

- Complexity of the format - creating and understanding MARC records often requires extensive knowledge.
- Designed to support card catalog access in libraries, not the generation of finding aids for archives.
- Lack of hierarchy - MARC generally lacks the tools to represent the inherent hierarchy in archival collections.
- Narrow domain - unlike EAD, MARC is not an application of a more general technology like XML, but an application specific format.
- A primary resource consideration in using a traditional MARC cataloging scheme are that it will require professionals or paraprofessionals with specialized cataloging knowledge. It will also require the use of specialized cataloging software to create and maintain the records.

4.2 EAD

Encoded Archival Description (EAD) is a Document Type Definition (DTD) that presents a standard for encoding archival finding aids in Extensible Markup Language (XML). XML, despite its name, is not a markup language; it is an

accepted standard for defining markup languages. A DTD is one way to define such a language: it unambiguously specifies in a machine readable way the structure of a document including what content elements can be included, where they can be included and contained, their names, and other information. The EAD standard is maintained by the Network Development and MARC Standards Office of the Library of Congress in partnership with the Society of American Archivists. (Library of Congress.)

Two important aspects of EAD's historical context are worth noting. First, its origins lie in the archival community. Unlike MARC, EAD was developed specifically for archives by archivists. Second, it has direct connections to previous work in worldwide work in codifying archival standards. EAD does not exist in isolation. Prior to EAD's development in 1995, the International Council on Archives' Ad Hoc Commission on descriptive Standards adopted the General International Standard Archival Description format - ISAD(G). (Fox 2001) ISAD(G) is a conceptual model that clearly explains what it is that archival description needs to capture, and EAD is an implementation of that model in an encoding format.

What is less clear is EAD's relationship to content standards specifying exactly what the generally described elements and structures should contain and how to present them. Local practices hamper interoperability, but are one option. APPM can be used, as well as the Canadian equivalent Rules for Archival Description (RAD), and there is ongoing discussion on unifying them. (Fox 2001) In addition to encoding standards, choosing controlled vocabularies is another important content standards issue that is not clearly solved by EAD.

Since EAD is XML, it can leverage existing and continuing work in general XML technologies, like XSLT, a well-documented specification to transform XML documents into other formats. Although not a verifiable quantitative fact, it is likely the case that XML has a large mindshare of technical developers and is a growing community, which is likely not the case for MARC.

Strengths of EAD include:

- Hierarchical nature of EAD documents reflects multilevel description practices central to archives. (Haworth 2001)
- Designed specifically for the creation of finding aids
- EAD is relatively easy to convert to other formats, especially web sites.
- Extensible and customizable, especially in comparison to databases (Clough 1998)
- Increasingly accepted as the future of archival description, many institutions are adopting it and discussing positive results of their work (EAD Help Pages 2004)

Weaknesses of EAD include:

- EAD is generally stored in text files - for large projects huge text files are not easy to create, edit, and manage in comparison to database systems.

(XML databases address this issue but are an emerging technology not ready for large scale adoption.)

- Although many tools exist for editing XML documents, most are still complicated and do not integrate authority control like MARC cataloging systems
- Content standards for EAD are not agreed upon and may hamper interoperability
- Designed for finding aids - not as a purely records-oriented data standard makes it inherently more flexible, but also more complex and with variation in implementations across institutions
- Popular toolsets are difficult to use, have serious usability limitations (Prom 2002)

Resource considerations with an EAD project include finding the technical expertise to put together an EAD software suite suitable for editing, validating, storing, and transforming EAD documents. Much XML software is available, and much of it is free, but usability may be just as big of an issue as with proprietary MARC-based systems. Additional work may be necessary in choosing what content standards to use and how to integrate them with the system.

5 Recommendations

5.1 Crosswalks

One possible solution would be to pick one format, EAD or MARC use that for cataloging and storage, and develop what is termed a "crosswalk" to the other format, thereby gaining many of the advantages of both. Crosswalks refer to a mapping of one metadata format, like MARC or EAD, to another. (Greenberg 2002) Crosswalks are very useful, but they are not a perfect solution. There is almost inevitable data loss as the mappings may not capture everything. Different standards represent things at different levels of granularity - in this particular domain the problem of MARC subfields can be very tricky, as well as the multiple levels of description in EAD.

While crosswalks are important, and the work done in creating them should be examined in light of the proposed solution, we recommend against the strategy of developing a MARC or EAD system and relying on a crosswalk to the other in light of other more flexible solutions.

5.2 Databases designed for interoperability

Databases are software systems that manage vast amounts of data for storage, updating, and retrieval. Databases are robust, proven technology with a strong theoretical underpinning from the entity relationship model.

The conceptual model and the content standards matter as much if not more than the encoding formats. How that information is encoded is important, but it is only one step in a long process. (Maler 1995) The important thing is to understand the information needs of the project, and design a system that captures that. Therefore the recommendation is to develop a clearly articulated local standards practice before any cataloging begins. This should have the following overall steps:

1. Clearly document current and future project goals, specifically in terms of output format: including printed finding aids, web sites, federated archives initiatives participation.

2. Design the conceptual model: begin with the ISAD(G) model, and clearly document any areas in which it will not be sufficient for the catalogs, indexes, and finding aids for this collection.

3. Design the database: create a database scheme reflecting the model but specifically with the design goal that the mapping from this database to both EAD and MARC, documented, clear and without information loss.

4. Choose content standards: recommend using APPM where appropriate as a content standard to better ensure interoperability, as well as designing an authority control for the collection, and using other established thesaurus or controlled vocabularies.

5. Choose or design components, tools, and systems for the cataloging in cooperation with actual catalogers, and iterate the design of the system until it functions for the catalogers sufficiently. Recommendation is to use open-source tools to decrease the cost of software, and using that saved money on labor to customize systems or build needed in-house tools.

6. Catalog the scanned collection using the system and practices outlined above

This database strategy is similar to the used for the Robert J. Honeyman Collection, one of the first large scale uses of EAD for a visual collection, and was quite successful. (Elings 1998)

Advantages of a database strategy:

- Gain many of the advantages of MARC and EAD since they can be created easily from the database

- Not locked in to any one format, room for growth, incorporation of new formats into the system

- Databases are designed for reliable storage and data integrity

- Can be integrated with many different software packages and systems for updating and creating records (desktop applications, web sites)

- Ability to choose open-source toolkits decreases costs, ensures long term viability of system

Disadvantages and limitations include:

- Converting database records to EAD and MARC without loss or difficulty is a non-trivial problem
- More difficult to represent multi-level description and hierarchy than in XML
- Less flexible during record creation as EAD
- Added abstraction layer of the database adds complexity to the process
- New encoding formats will require new mappings
- More upfront work in designing the standard and system to create records
- Using a homegrown system means little to no vendor or company support

One of the biggest tradeoffs in using a database solution is that it focuses more on the data-centric nature of records, rather than the document-centric nature of finding aids as EAD does. While this is an issue, expanding the database to include the information that would have been in finding aids but not catalog records helps to bring all the information about the collection to a single place has significant advantages, as discussed below.

Although the disadvantages of this solution are outlined above, a well thought-out design phase coupled with iterative designs and testing should alleviate most of them. Furthermore, the advantages in terms of flexibility, reliability, and opportunities to create a quality open toolkit are substantial. As a government funded, public institution, the work done on this project should be looked at not just in terms of its tangible records output, but also with an eye to developing a reusable open system that can be shared with other institutions.

5.3 Output Formats

One of the important design goals of this project is to be able to output different formats. By moving the cataloging information to a clearly documented database format, the possibilities for single-source publishing are much better. In addition to being able to output MARC records for OPAC and library catalog integration, and EAD for its strengths as an interchange format, other output format including XHTML for web pages, PDF for high quality paper documents, and Dublin Core metadata for web resource discovery are all possible. Since MARC and EAD are easily created by this database, existing work in converting MARC and EAD to other formats can be leveraged where appropriate, cutting down on duplicative work.

An important aspect here is that by having a single database, rather than only data-centric MARC records or structured EAD XML files, we can create

both EAD and MARC records, and there is a unified source for creating both catalog records (in MARC or other formats) and descriptive finding aids (in EAD or other formats.) Updates to the database can update both the finding aids and the catalogs, eliminating duplicate work and rekeying information.

6 Project Scope, Costs, Timeline

As evidenced by the preponderance of design and conceptual preliminaries in this project plan, there will be high upfront costs in approaching the project in this manner. These upfront costs will be substantial in time and labor if a participatory, iterative design process is used to create the cataloging tools. In addition to programmer time and labor, catalogers will have to take time out from cataloging to help design and test the system. However, these upfront costs are likely worthwhile as they will create a cataloging system that fits the needs of this project and will not require extensive changes during the cataloging process creating catastrophic cost and time increases.

A projected timeline for the development of the system: 1 month - conceptual design, modeling, choosing content standards
2 months - implementation of system, integrating components
1 month - testing, redesigning components
1 month - small pilot cataloging experiment, final redesigns

This could likely be completed by 2-3 programmers in-house or outside consultants at a cost of $20-30k on labor. Open source tools should be used and will decrease software costs. Proprietary systems are another option, and will have the advantage of company support, but are not recommended as customization and proper systems design are integral to this project.

The actual cataloging costs will depend on the speed of cataloging. The participatory design process will likely increase efficiency of cataloging, but it is still a time-consuming and laborious process. Assuming a cataloger processes 5-10 records an hour, and is paid $10 an hour, the cataloging of 25,000 photographs will cost at least $25,000 and up to $50,000. Scanning can proceed independently of the cataloging, but cost will likely be at least $10,000 depending on the speed and cost of labor. This is not counting initial upfront costs of additional computers, scanners, or other equipment.

While the scope of this project is the Scurlock collection, a well designed system that performs adequately here can be used for other similar cataloging tasks in the institution, and innovative solutions found can be shared throughout the archival community, helping the field as a whole.

7 Bibliography

1. Clough, Matthew H. “A Question of Access.” Archives and Museum Informatics 12(1998): 293-298.

2. "EAD Help Pages: Implementation Overview." August 18, 2003. EAD Round Table of the Society of American Archivists. [http://jefferson.village.virginia.edu/ead/sitesann-index.html] Sept 18 2004.

3. "Encoded Archival Description (EAD): Official EAD Version 2002 Web Site." 12 Aug. 2004. The Library of Congress. 18 Sept. 2004 [http://www.loc.gov/ead/].

4. Elings, Mary W. and Eva Garcelon. "The Robert Honeyman Jr. Collection Digital Archive: EAD and the Use of Library and Museum Descriptive Standards." Archives and Museum Informatics 12 (1998): 205-219.

5. Fox, Michael J. "Stargazing: Locating EAD in the Descriptive Firmament." Journal of Internet Cataloging 4.3/4(2001): 61-74.

6. Greenberg, Jane. "Metadata and the World Wide Web." in Allen Kent and Carolyn M. Hall (Eds.) Encyclopedia of Library and Information Science. Marcel Dekker, New York, 2002, pp. 244-261.

7. Haworth, Kent M. "Archival Description: Content and Context in Search of Structure." Journal of Internet Cataloging 4.3/4(2001): 7-26.

8. Hensen, Steven L. "Archival Cataloging and the Internet: The Implication and Impact of EAD." Journal of Internet Cataloging 4.3/4(2001): 75-95.

9. Maler, Eve and Jeanne El Andaloussi. Developing SGML DTDs: From Text to Model to Markup. Prentice Hall PTR; 1st edition (1995)

10. Prom, Chris. "The Ead Cookbook: A Survey and Usability Study." American Archivist 64 no. 2 (2002.)