Against the Grain

February 2007

Balancing the Needs of Producers and Managers of Digital Assets through Extendible Metadata Normalization

Arwen Hutt
University of California, San Diego, ahutt@ucsd.edu

Trish Rose-Sandler
University of California, San Diego, trose@ucsd.edu

Bradley D. Westbrook
University of California, San Diego, bradw@library.ucsd.edu

Follow this and additional works at: https://docs.lib.purdue.edu/atg

Part of the Library and Information Science Commons

Balancing the Needs of Producers and Managers of Digital Assets through Extensible Metadata Normalization

by **Arwen Hutt** (Metadata Librarian, University of CA, San Diego) <ahutt@ucsd.edu>

and **Trish Rose-Sandler** (Metadata Librarian, University of CA, San Diego) <trose@ucsd.edu>

and **Bradley D. Westbrook** (Digital Archivist & Metadata Librarian, University of CA, San Diego) <bradw@library.ucsd.edu>

Introduction

The **UCSD Libraries' Digital Library Program** was formally established in 2001. Like such programs at many other research university libraries, **UCSD**'s digital library program had its inception in a number of disparate digital library projects that took place during the 1990s and early 2000s. Some of these projects, such as **The Visual Front: Posters of the Spanish Civil War** (*http://orpheus.ucsd. edu/speccoll/visfront/index.html*), were digital exhibits designed to expose selected special collections and archival materials to a wider audience. Others, such as the **Digital Audio Reserves** (*http://orpheus.ucsd.edu/music/*), were designed to extend the libraries' reserve services by providing more accessible digital copies of the audio materials. Yet, others were proof of concepts or motivated by external collaborations, such as the digitization of art slides to populate the **ARTstor** database (*http://orpheus.ucsd.edu/slide/arts.html*) or **California Explores the Oceans** (*http://ceo. ucsd.edu/expeditions/*).

Paradoxically, but not surprisingly, the one common characteristic shared among these disparate projects was that they were created with local project needs in mind, without regard for interoperability with other library projects or long term preservation. This was especially true for the metadata associated with these projects. Consistent and standardized metadata is crucial for interoperability between digital collections and for the long term management and preservation of digital objects. In particular, metadata that allows the preservation function to be carried out, commonly called preservation metadata, was either inconsistent or incomplete. Preservation metadata includes, to a lesser or greater degree, descriptive, administrative (technical, provenance and rights) and structural metadata. How much preservation metadata is required depends on the preservation functions of the custodial repository.[1] As for the **UCSD** projects, certain descriptive metadata elements were not always labeled the same or recorded according to the same conventions. In addition, the technical metadata for authenticating files was often lacking, or in vendor constructed spreadsheets. Rights metadata was uniformly

lacking, but not inexistent; it was often to be found in paper control files describing the collection. Structural metadata, where present, was encoded as parts of file names.

Consequently, when the **UCSD Libraries** decided to build a **Digital Asset Management System (DAMS)** in order to bring its digital assets under a common management and access framework, several critical questions had to be addressed. Chief among these was what to do with the metadata for the legacy assets, some of which existed in the libraries' **MARC** information library system, while other metadata records were stored in local databases or spreadsheets. Should the libraries adopt a lossless process and import the metadata as it existed and then establish post-import procedures to make the metadata interoperate? Or should the libraries stipulate metadata standards for the **DAMS** and then normalize legacy metadata to them upon import, recognizing, of course, that some legacy metadata might be lost or transformed in a way not always desirable to the content producer?

Extensible Normalization

The libraries' **Metadata Analysis & Specification Unit (MASU)**, comprised of the three authors listed above, proposed an approach that might be best called "extensible normalization." The unit decided normalization was the best approach to centralizing the libraries' legacy data, ca. 300,000 digital assets and associated metadata records.

Normalization

Normalization is a formal analytical process by which various metadata formats are standardized to a pre-selected metadata standard, e.g., the local **Excel** spreadsheet of artists' names is standardized to the **Union List of Artists' Names (ULAN)**. This process involves direct element to element mapping as well as more complex data relationships and content processing procedures. Normalization ensures that the basic metadata requirements for achieving interoperability and efficient management across all objects are satisfied at the outset. It ensures that the metadata formats stored adhere to community standards, thereby making the data content easier to use either in other repository environments, for example the **Online Archive of California**, or for aggregation via the **Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)**.

Normalizing data on import also lessens the cost of data centralization. The number of schemas necessary in the **DAMS** is minimized, thereby reducing the complexity of the system required to manage and

maintain the data store. This saves cost in the initial stages of system development and allows technical resources to be focused on the building of a robust system. Constraining the data elements in the **DAMS** minimizes the need to modify management, reporting and exporting processes, thereby reducing the overhead involved in managing data over time.

One of the most important advantages, within a distributed organizational structure such as at the **UCSD Libraries**, is that normalization does not necessarily place the burden of work on the content producers. With this approach they are under no obligation to alter their metadata production processes in order to conform to the **DAMS** metadata standards. In addition this approach may be especially beneficial to organizations that have minimal IT support for digital library development.

Extensibility

At the same time, we recognized the metadata standards initially selected would be unlikely to meet the needs of all communities to be eventually represented and served by the **DAMS**. Indeed, our Art and Architecture Library, which currently uses **MARC** for expressing metadata for its digital images, has been contemplating how to produce **VRA Core** compliant metadata records to take advantage of the **VRA** hierarchical data model and as well as its content and context descriptors. From this single fact, it was clear our approach had to be extensible, open to the eventual addition of other metadata standards where useful for the accurate and successful representation and delivery of resources.

We believe our approach of extensible normalization strikes an effective balance between the need for centralized, cost-effective, and interoperable data management and our traditionally decentralized administrative organization. At **UCSD Library**, departments have traditionally enjoyed a certain amount of organizational autonomy, not only in respect to selecting resources but also in respect to organizing, describing and providing access to those resources. This has been especially true for digital materials.

Data Import Processes
Data Inventory

To understand **MASU**'s approach to normalization, it may be helpful to describe our processes for preparing data for import. The first stage was to inventory what digital assets were available for importing into the **DAMS**. As stipulated in the Content Policy for **UCSD Libraries' Digital Asset Management System**, digital content imported into
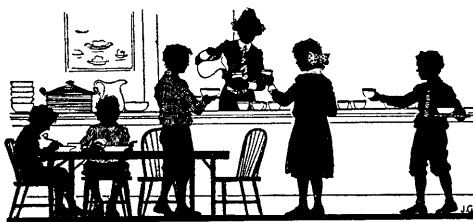
the **DAMS** is restricted to material created by the **UCSD Libraries'** content producers and selected by them for inclusion in the **DAMS**.[2] This excluded materials such as commercially acquired electronic resources. Content producers were contacted and interviewed in order to help assess which resources complied with the **DAMS** content policy.

Results of the inventory were written up in a report for the library entitled the *Digital Collections Assessment Report*. The report sought to characterize the producers' content files[3] and accompanying metadata. **MASU** quantified not only the number of content files that each producer created but also the number and type of digital objects they represent. A digital object is "an entity in which one or more content files and their corresponding metadata are united, physically and/or logically, through the use of a digital wrapper." A digital wrapper is a structured text file, often **XML**, which binds together content files and metadata and specifies the logical relationship among them.[4] In the library community the **Metadata Encoding and Transmission Standard (METS)**[5] is the most commonly used type of digital wrapper, whereas the multimedia community has shown greater interest in the **MPEG-21** standard.

There are two basic types of digital objects: simple and complex. Simple digital objects are "comprised of a single content file (and its format variants or derivatives) and the metadata for that file." An example of a simple digital object is a photograph represented by a high quality master digital image and one, or more, smaller delivery quality digital images. Complex digital objects include "two or more content files (and their format variants or derivatives) and corresponding metadata. The content files are related as parts of a whole and are sequenced logically, such as pages." An example of a complex digital object is a multipage pamphlet where each page is represented by a separate digital image.[6]

The inventory also characterized metadata by type including: descriptive; administrative and structural. The inventory exposed not only the wide variety of metadata that had been produced but also highlighted areas where metadata was lacking and would need to be added in order to provide long-term preservation.

Establishing Metadata Targets

The second stage of work involved using the data gathered from the inventory to help **MASU** determine which community metadata standards the **UCSD Libraries** should adopt to serve as normalization targets for descriptive, administrative and structural metadata, as well as how those standards would be adapted to the local environment. **METS** was chosen as a digital wrapper for binding together metadata and associated content files because:

- It is the leading digital wrapper format within the digital library community and has growing institutional support.
- It officially supports academic metadata standards such as **DC**, **MODS** and MARCXML.
- It allows the inclusion of different metadata schemas to describe different facets of an object (descriptive metadata vs. administrative metadata) and different representations of an object (**DC** vs. **VRA**).

MASU chose **Metadata Object Description Schema (MODS)**[7] to serve as the common descriptive data standard within the **DAMS**. **MODS** was chosen because:

- It allows rich resource description without an overwhelming element set or overly complex expression of structure.
- Its basis in **MARC** makes it easier to transform the variety of **MARC** based data in the **UCSD Libraries**.
- Its origin in description of bibliographic materials means many of the concepts are familiar within a library environment, but this focus isn't so strong that non-bibliographic materials can not be described accurately.
- It has strong community support and is maintained by the **Library of Congress**.

PREservation Metadata Implementation Strategies, or **PREMIS**[8] for short, is a data dictionary that identifies core preservation metadata elements and was chosen for non-format specific administrative metadata. It is expected that expression of a set of common technical metadata elements across file formats, will improve the efficiency of preservation management within the **DAMS**. The only format specific technical metadata schema that we have currently adopted is the **NISO Technical Metadata** for **Digital Still Images**, expressible in **XML** using the **MIX** schema.[9] We will add other community-endorsed format specific standards as they are needed.

An essential part of the process of establishing metadata targets is adapting the selected metadata standards to the local data environment. For instance we struggled early on with the problem of whether to anchor our **MODS** descriptive metadata within the digital manifestation or within the intellectual work itself. This decision impacted how we interpreted and used many elements including publisher and type of resource. A variety of resources exist to support this decision making process. These include schema documentation itself, support listservs such as the **PREMIS Implementors Group Forum**[10] and the **METS listserv**, and best practice guidelines such as the *Digital Library Federation's MODS Guidelines*[11] and *CDL's Guidelines for Digital Objects*.[12]

Object Specification

The third stage of data preparation is writing the object specifications. Generally an object specification stipulates what kinds of content files and metadata types are permissible, what metadata elements are required, how legacy metadata is to be treated and how the content files are to be referenced by metadata. **MASU**'s object specifications consist of four parts. The most basic and abstract part is the **METS** profile. A **METS** profile stipulates the basic requirements (metadata and file format) for a particular class of digital objects, for instance, maps or electronic theses and dissertations. The second part is the source to target mapping, which serves to indicate how the legacy metadata created by a particular content provider is to be treated. The third is a target from source mapping, which serves as a blueprint for the object to be assembled from the content provider's legacy metadata and content files. The fourth part is a hand assembled object that has been validated.

Content producers also play a role in formulating the object specifications for their materials. They provide input on what data must be mapped into the **DAMS** and what can be left behind. Of the data that is mapped into the **DAMS**, the content producers may stipulate how that data is to appear, may help to disambiguate ambiguous content, and resolve gaps between source and target metadata. As an illustration, our work with the **Scripps Institute of Oceanography Archives** identified several fields in their original database which were utilized solely for internal workflow management, and so we were able to eliminate these from the mapped data with no loss of functionality. We also learned that the content of one field, which had been created for a particular project, could be useful if the value "oceanography" could be discarded. This term was applied when more specific categories did not adequately describe a resource, and although relevant within the context of the original project, it was likely to be misleading in its new context. Without the input of content producers issues like these would be more difficult to handle with confidence, and in some cases may be missed altogether.

Object Assembly

The final stage before data import is the iterative process of assembling the objects. An assembly package is handed off to the Information Technology Department (ITD), which includes the object specification documentation as well as detailed information on where the digital objects and metadata are located and how to identify them. An Extraction Transformation and Load (ETL) specialist within ITD then writes and executes code for building the objects. The assembled objects are then returned to **MASU** for quality assurance review, before final assembly and uploading of the objects to the **DAMS**.

Future Questions

In addition to the many advantages of this approach, **MASU** has also considered two potential problems. Our normalization approach may be better suited to the conversion of legacy data than to continuous real time data production. Integrating dynamic, developing collections into this process is possible but will likely result in unacceptable inefficiencies. Our normalization process is based on batch processing of legacy digital assets. It would not apply well to assets created in real time; their through-put would be delayed considerably. Inefficiencies could also be introduced when content producers change their local metadata or content standards in such a way that requires a modification of the object specifications.

Although this may be feasible in the short term, expecting this in the long term is neither realistic nor consistent with the overall policy of distributed local control.

Another potential difficulty is the increasing asynchronicity between the content producer's metadata in their local production/delivery environment and the transformed metadata that represents their objects within the **DAMS**. This asynchronicity will grow as content producers modify their local data. The problem is relatively minor as long as the objectives of the local data environment and the **DAMS** are different, that is to say the local metadata serves access whereas the **DAMS** metadata serves preservation and overall collection metadata. But the problem is greatly exacerbated should the **DAMS** become an access instrument as well. In such an event, efforts will need to be made to insure synchronicity between the local database and the **DAMS** or, more radically, the local database will be subsumed into the **DAMS**.
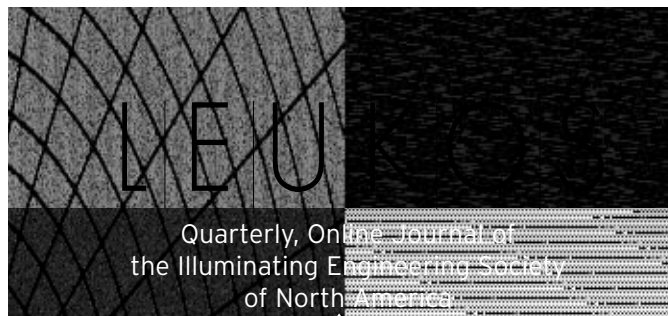
Conclusion

In conclusion, this approach to the centralization of digital assets provides many immediate and long term benefits. An immediate benefit is the attention it gives local collections and the needs of content producers, and the rapidity by which all the libraries digital assets are brought under common preservation management. The process of working with content producers helps **MASU** to better understand their materials, users, and expectations and consequently to define more accurate object specifications.

Endnotes

1. As stated in the *PREMIS Data Dictionary for Preservation Metadata*, preservation metadata is defined "as the information a repository uses to support the digital preservation process. Specifically, the group looked at metadata supporting the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context. Preservation metadata thus spans a number of the categories typically used to differentiate types of metadata: administrative (including rights and permissions), technical, and structural. Particular attention was paid to the documentation of digital provenance (the history of an object) and to the documentation of relationships, especially relationships among different objects within the preservation repository" See page ix in *http://www.oclc.org/research/projects/pmwg/premis-final.pdf* [accessed 11 Jan 2007].

2. In keeping with the goals of balancing the content producer's needs with the needs of the **DAMS**, it was determined there would be no restriction on the kinds of file formats that could be imported into the **DAMS**.

3. Content file is "a file that is either born digitally or produced using various kinds of capture application software. Audio, image, text, and video are the basic kinds of content files." All definitions in this article that are in quotations were taken from the *CDL Glossary http://www.cdlib.org/inside/diglib/glossary/* [accessed 29 Nov 2006].

4. See "digital wrapper" in the *CDL Glossary http://www.cdlib.org/inside/diglib/glossary/?field=term&query=digital+wrapper&action=search* [accessed 11 Jan 2007].

5. **METS Encoding and Transmission Standard** *http://www.loc.gov/standards/mets/* [accessed 29 Nov 2006].

6. See definitions for "complex digital object" in the "Digital Objects" section of the *CDL Glossary* (*http://www.cdlib.org:8081/inside/diglib/glossary*).

7. *MODS User Guidelines Version 3 http://www.loc.gov/standards/mods/v3/mods-userguide.html* [accessed 29 Nov 2006].

8. **PREMIS (PREservation Metadata: Implementation Strategies) Working Group** *http://www.oclc.org/research/projects/pmwg/* [accessed 29 Nov 2006].

9. **NISO** Metadata for Images in XML Schema *http://www.loc.gov/standards/mix/* [accessed 29 Nov 2006].

10. **PREMIS Implementors' Group Forum** *http://www.loc.gov/standards/premis/pig.html* [accessed 5 Dec 2006].

11. *Digital Library Federation MODS Implementation Guidelines for Cultural Heritage Materials* http://www.diglib.org/aquifer/DLF_MODS_ImpGuidelines_ver4.pdf [accessed 29 Nov 2006].

12. CDL's Guidelines for Digital Objects http://www.cdlib.org/inside/diglib/guidelines/ [accessed 29 Nov 2006].

of money for public relations campaigns by all stakeholders, including, for example, the **PLoS** advertising and marketing budget that was close to a half a million dollars in 2004. **PLoS** is not the only group that has launched such campaigns; **SPARC** has been very aggressive, and now the **Association of American Publishers** has retained a public relations guru.

Thousands and thousands of dollars are being expended on the pro-con open access debate, and yet it has not been fully examined from a fiduciary point of view.

Without a sound fiduciary model that is sustainable, all the rest is an exercise of eloquent (and very repetitious) prose. And wasted money.

We do not know if the money for sustainability and affordability is assured. Who is going to demand that answer? Until we have long standing evidence of sustainable and affordable models, we have to be absolutely sure that ideological fervor does not overtake the realities of what all this will really cost, and, please…. Repeat after me, where will this money come from? And for how long?

Does the subscription system have flaws? Indeed it does. Should publishers and librarians still try to create a better system together while we grapple with the unknown? Indeed we should.

Beware of unintended consequences. It is well to keep in mind the phrase "Don't Throw the Baby out with the Bath Water." Credited to the first written occurrence in the satirical book, *Narrenbeschwörung (1512)*, by **Thomas Murner** (1475-1537), a chapter is entitled such: it is a treatise on fools who by trying to rid themselves of a bad thing succeed in destroying whatever good there was as well.

Well said. And very good advice.

And remember to send for your application to law school. 🌱

A longer term benefit of this approach is the development of understanding and familiarity between **MASU** staff and content producers. It is hoped these relationships will increase their comfort with approaching **MASU** for future assistance or advice regarding metadata or cataloging. Moreover, it provides a tested model for working with content providers outside the library, say the engineering faculty, who want to contribute materials to the **DAMS** for safeguarding.

MASU is confident our extensible normalization approach meets the needs of aggregating legacy data while remaining flexible enough to evolve along with the changing needs of the **DAMS** and the **UCSD Libraries**. 🌱

ATG Special Report — Cataloging eBooks: an Overview of Issues and Challenges

by **Kristin E. Martin** (Electronic Resources Cataloger, Davis Library, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514-8890; Phone: 919-962-0153; Fax: 919-962-4450) <kmarti@email.unc.edu>

Web-based eBooks have become popular with a wide variety of library users and are an increasingly important part of libraries' collections. eBook content now encompasses databases of retrospective eBooks (such as *Early English Books Online* or *Literature Online*), aggregated packages of relatively current content from multiple publishers provided by an eBook vendor (such as **NetLibrary** or **ebrary**), and titles offered directly from the publishers (such as **Springer** and **Elsevier**). As the volume of eBook content grows, libraries are grappling with how to integrate this content into their online catalogs. Librarians trying to provide title-level catalog access to their eBook collections must answer multiple questions to determine optimal workflow. Questions include:

- Where will the record come from?
- Can the eBook records be processed in batch?
- Should electronic holdings be placed on the same record as print holdings?
- What changes will need to be made to vendor-supplied records?
- How can the records remain accurate as titles are added and subtracted to eBook collections?
- Should holdings be added to **OCLC**? Why or why not?

At the **University of North Carolina at Chapel Hill University Libraries** we have been analyzing the issues raised by these questions to figure out how to provide the best access to our growing number of eBook collections. This article does not purport to be able to answer all of those questions, but rather introduces them as a series of topics that librarians will need to address when adding eBook records to their catalogs.

Although many eBook collections offer their own search mechanisms, having individual title records for eBooks in the OPAC provides library users with a single discovery tool for eBook titles across all collections and allows users to simultaneously view the library's print and electronic holdings. Initial studies of eBook use, mainly looking at **NetLibrary** content, have demonstrated the importance of catalog records in enhancing use to electronic books (for example see **Dillon** 2001; **Gibbs** 2001; **Langston** 2003). In a particularly dramatic example at the **University of Rochester**, the use of the **NetLibrary** eBooks increased by 755 percent when comparing use in the five months before and after loading the catalog records (**University of Rochester Libraries** 2001). Later studies of eBook usage have taken title-level catalog records for granted, when comparing usage of print and electronic counterparts (**Christianson** and **Aucoin** 2005; **Littman** and **Connaway** 2004).

Despite the preponderance of evidence supporting the need for access to eBooks through the catalog, many libraries have been quicker to purchase eBooks than to provide title-level access through the OPAC. Several issues have contributed to this delay in cataloging. Acquisitions and cataloging workflows have been developed around the processing of physical items, generally on a title-by-title basis, while eBooks are intangible objects that have frequently been made available in large collections that could overwhelm a cataloging department. Staff may still have a "print is primary" mindset, and view electronic resources as supplementary, rather than as a core part of the library's collection. Additionally, eBooks may only be available on subscription, rather than owned, and titles may be swapped in and out as new material becomes available in large collections. Finally, cataloging standards for electronic resources have been subject to multiple revisions, making libraries reluctant to spend time and resources creating catalog records that will need to be updated.

Fortunately, as eBooks have become more widespread, so has the availability of **MARC** records for individual titles, frequently from the vendor. One of the first questions librarians must consider is whether to use vendor-supplied records for eBook collections. Records may be free with the purchase of the resource, available for a fee through **OCLC's Collection Sets**, or available for purchase separately from the vendor, with price and quality of records varying widely. These vendor-supplied records free the library from having to provide title-by-title cataloging, and may be loaded quickly into the catalog; however, there is still work to be done at the library's end.

Librarians must scrutinize the records carefully for quality and ensure the records correctly represent the titles the library purchased. Given the size of some eBook collections, it may not be possible to examine each record, but it is important to at least spot check records or to examine a selective sample for quality and accuracy. To date, vendor records have typically treated eBooks as electronic