The International Journal of Digital Curation
Issue 2, Volume 3 | 2008

Applying the Digital Curation Lessons Learned from American Memory

Liz Madden,
Office of Strategic Initiatives,
Library of Congress

November 2008

Summary

American Memory, launched in 1995, was the Library of Congress's debut web presentation and the primary product of the National Digital Library Program. More than 700,000 described digital items in 90 collections were added to American Memory in those first five years, including content from 23 external organizations. These materials were digitized, assembled and presented without tools designed specifically for the work and before the development of approved standards for the creation, presentation, or exchange of digital content. Valuable lessons about all levels of digital curation emerged from this early foray into digital library work, and many of the issues have persisted into current digital library efforts at the Library. This article focuses primarily on lessons learned about the conceptualization, creation, receipt, and preservation actions for digital content. It describes how strategies developed early on to manage the diverse and heterogeneous digital content helped inform later practices and were applied to legacy data in an effort to ensure their sustainability, flexibility and shareability into the future.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Digital Curation Lessons Learned

Background of American Memory Data

The Library of Congress created the American Memory (AM) presentation in 1995 to provide public access to digitized versions of Library collections. During the first phase of AM production from 1995 to 2000, more than 700,000 described items from 90 collections were added to the website. The materials came both from the Library of Congress collections and from awardees of the Library of Congress/Ameritech Digital Library Competition, a three-year program aimed at enabling U.S. libraries, archives, museums, and historical societies to digitize their own collections of Americana and provide access to them through the AM website. Library of Congress staff who prepared the collection contents for digitization, oversaw the scanning, assessed the quality of the digital text and images, and transformed the data for use with the presentation programming, were also responsible for creating or enhancing the descriptive metadata for the digital objects.

Among the physical materials digitized for the AM presentation were photographs, negatives, maps, atlases, manuscripts in paper form and on microfilm, bound printed books, sound recordings, motion pictures from film and paper prints, posters, pamphlets, sheet music, and even a physical collection of musical instruments. The sources of descriptive metadata ranged from digital forms such as MARC records (in binary form because this was before XML), non-EAD (Encoded Archival Description) finding aids in word-processing formats, and fielded proprietary databases, to non-digital forms such as printed tables, lists of items contained in card catalogs, manuscript collection boxes and microfilm. Many objects had no descriptions at all.

AM was one of the first large-scale digital library presentations on the web. All of the work to create an online virtual presentation of the diverse physical materials and descriptions was accomplished without production tools or standards specifically designed for digital library content creation, description, exchange, or display. The staff who produced the digital content learned through trial and error what worked and what did not for large-scale production. They adjusted processes and practices along the way to improve productivity and efficiency based on earlier experience.

Early Lessons Learned

A more detailed description of the major lessons learned was presented at the DigCCurr2007 conference in Chapel Hill, NC, in April 2007 (Madden, 2007). A brief summary of those early lessons follows.

Know thy data.[1]

An understanding of the intellectual nature, intended use, and the relationship between the intellectual and the digital is critical to the preservation and presentation of digital content. An understanding of the intellectual value or nature of content helps inform decisions about digitization and presentation. Knowledge of how the digital content is used or presented can aid in the development of schedules or policies for

[1] Martha Anderson is the author of the first five rules of "Everything we needed to know we learned from digitizing" upon which I draw in the following paragraphs.

backup, archiving, or integrity checking. Identifying where errors in descriptive metadata transformation have occurred in the past can help predict where future errors might be introduced to data of a similar nature.

Automation means letting machines do the work.

Automation reduces human error and provides for repeatable and consistent processes. Automated processes are more likely than human processes to make mistakes in a consistent and repeatable manner, which can make it easier to predict, identify and correct errors. Automation can act as a catalyst for developing standards because it relies on consistency of practice and data. Any automation counts. Many data irregularities are introduced when source data—e.g., existing descriptive records—must be transformed from a seemingly unique existing format into a new format for use in an application or presentation. The temptation to do all the transformation manually for expediency can be very strong, but many apparently unique transformations will recur, and automations as simple as word processing macros or simple scripts using regular expressions for search-and-replace can save time and introduce a modicum of consistency. Documentation also saves time and headaches by relieving those doing the transformations of the burden of having to remember just exactly what it was they did the last time they saw a certain kind of data.

Exceptions to rules raise resource usage.

Exceptions can occur at any stage of digital content creation. They can come in the form of added descriptive record fields that must be accommodated in a display, or in extra lines of code written to treat one set of objects differently from other objects of the same type. Customization is also a kind of exception. It reduces the sustainability of data and the scalability of production by requiring extra or special processes to interact with a subset of data. Exceptions create data that diverge from regular practice and as a result become more difficult to sustain in the long term. Custom treatment often relies on institutional memory or documentation for long-term maintenance, both of which may be absent. In addition to this, solutions for newly identified application or data problems may not be applicable to processes that have been tailored to fit special cases. Standards themselves are also subject to customization or differences in interpretation. There is a cost when data created according to one standard enter an environment where the same standard was interpreted differently.

Interoperability requires compromise.

Contents must contain required elements and be structured in an expected way in order to be shared among multiple environments and to work in different presentations or applications. Upfront agreements about data are integral to collaborative digitization projects. These agreements ideally establish delivery schedules for data; ensure persistence of data structure for corrections or future deliveries or redeliveries; designate the party responsible for maintaining the "master" data; identify a minimum set of required data elements, and provide explicit details of any rights or permissions restrictions related to the display or other use of the content.

Simply choosing a descriptive metadata standard is insufficient because metadata standards are not data models and may not contain information necessary to integrate with an application or another environment. MARC is a well-supported standard, but even records that came in MARC format often had to be manipulated or enhanced to work in AM. If data are being transferred across different systems platforms, then file

naming rules might also be part of pre-exchange discussions to avoid problems related to the different handling of characters such as \, *, and $ across platforms.

Diversity must be recognized.

Heterogeneity of data is a fact of digital library work. Content producers have different needs for their data and applications, and they must create practices, routines, and data standards that serve their own work. Even users within a single community may have different needs for data based on what portion of the lifecycle they focus on. Recognizing when to enforce conformance to existing structures or practices and when to accommodate diversity is critical to creating data that can be used and sustained. Choosing or developing data models that provide for flexibility within predefined constraints and allow for automated reconfiguring of elements can be valuable in environments where data must serve more than one standard, community, or application.

Reduce, reuse, recycle.

The producers of AM recognized the value of well-organized and elemental data that could be reconfigured and reused to accommodate different portions of the digital production lifecycle. In addition to forming the basis for the display record in the presentation, descriptive record fields stored in relational databases could be concatenated or virtually manipulated without affecting the standardized data themselves. The more granular and descriptive-standard-agnostic the data, the more flexible they can be. Various AM project teams leveraged this flexibility to manage different aspects of their digital production. Some created scanning lists for the digitization providers; others generated letters and tracked mailings seeking permission to display content; still others generated metadata for TEI headers used in keyed SGML text transcriptions. Descriptive metadata creation can be time-consuming and expensive, and the ability to re-purpose the data for different stages of production or to accommodate different output standards may save time and resources in the long run.

Applying the Lessons Learned

Streamlining the Production

The only requirement in the first phase of AM was that the content had to have a title and an identifier. The title was used for discovery and description, and the identifier was used to locate the digital files associated with the title. Authors, creators, notes, dates, subjects, physical format, and place were desirable but optional. The digitized content was grouped into separate exhibits or "collections" that were thematic and/or representative of the underlying physical materials from which the digital content was created. Teams ranging in size from a single person to four or five people oversaw the production of a specific collection. These teams had subject or format expertise, and often sat within the division from which their collection content originated. Consequently, they often handled collections of materials of a similar physical and digital type, such as photographs, or maps, or books, or motion pictures.

In the early years AM display of a given collection was often tailored to the data in the descriptive records or the desired functionality for the presentation of the collection contents. When teams accustomed to working with one type of physical material, like photographs for example, encountered a collection with an unfamiliar type, such as a handwritten letter, the teams often invented their own custom way of

handling the new type rather than surveying the other projects to see what other teams had done with the same type of item. These customizations were based more on the physical format of the material than on the functionality of the presentation of that material. Thus books had one way of working, manuscripts another, and sheet music still another, even though all of them functioned similarly online as objects with multiple images that needed to display in sequence. Likewise, their descriptive records also varied according to what the teams were familiar with and how they wanted to showcase their materials.

The AMnonmarc storage record

Starting early on, AM presentation programming for descriptive record display was standard across all collections with underlying descriptive records in MARC format. These collections displayed the same fixed set of MARC fields with the same set of labels in the presentation. Any request to hide an existing MARC field or to show a new field had to gain consensus from other stakeholders with MARC displays because all would be affected by a change. For descriptive records that were not MARC, however, the number, labels and types of fields changed from collection to collection and required customizations to the presentation to accommodate. Nearly half of the collections in the AM presentation contained descriptive records that are not MARC. Display of this non-MARC descriptive data within the AM presentation varied greatly in the early years, depending on the source descriptive record format and the practices of the team producing the digital collections.

The teams liked the flexibility that came from working with records that were not MARC and developed new approaches to display that were not possible with the strict MARC programming. For example, manuscript collections applied a field-based sort-by-date browse feature to allow users to organize hit lists in date order. Folklife collections adjusted their records to illustrate the various roles played by content collectors and describers. Instead of "author" or "creator", the folklife collections could display with labels more fitting to folk materials, such as "Photographer" or "Interviewer." As the volume of AM content increased, it became impractical for the programmers to continue to accommodate all the customizations requested for digital object or descriptive record display, but the production teams had grown accustomed to a certain level of flexibility.

In the late 1990s a group convened to survey the practices and data of the different teams and to identify common elements upon which broader standards and practices could be applied. A user group of non-MARC record producers identified a common set of elements that could be used by all the non-MARC presentations. The first iteration of this AM non-MARC field set was integrated into a relational database tool that became the standard for new AM projects that did not have existing MARC records. The AM non-MARC storage record, or AMnonmarc, incorporated the flexible functionality that the teams had enjoyed in the past in a standard way that could be treated uniformly by the programming. Thus date sort and different roles could be included in the display with less burden on the programmers than before. The number of required fields increased as well. By 2001, all new AM collections without MARC records used this AMnonmarc record as an alternative to a custom field set.

Development of the AM Cookbook

A concurrent effort among the AM production teams also surveyed the existing data models and presentation functionality for objects in all the collections in AM. A basic set of data models was identified and documented in the AM "Cookbook of Collection Assembly." This cookbook provided project teams with the technical specifications for creating data according to specific models (a page-turner with text, a still image, a contact sheet, a large-format compressed image, etc.). The cookbook attempted to focus only on the functionality of the object in a web presentation and not on the source format of the physical material. For instance, simple, single still images were considered a single model regardless of whether they depicted photographs, negatives, small posters, or any other single image pictorial object. The "page-turner" model accommodated any multi-image objects whose images had to display in sequence, and so on. This helped curb the problems caused by collection or content-type customization.

Preserving the Legacy

American Memory Legacy Data

The AMnonmarc storage record and cookbook provided a standard tool, guidelines and data model specifications for the AM presentation. This streamlined and rationalized production of new AM content, but the problem of the early data remained. Hundreds of thousands of custom descriptive records and objects were still being served through the aging AM presentation. Due to a goal of putting a large amount of content online within five years, the production rate in the late 1990s was intense, and the project teams did not have time to document all the small decisions about descriptive metadata, display or functionality, let alone to normalize the existing sets to conform to evolving models. By the early 2000s, AM users continued to identify problems or errors needing correction in the old collections, but finding the record sets and identifying how to re-output the data had already become somewhat of a treasure hunt.

Characteristics of the Legacy Data

AM legacy data was inconsistent across collections, highly customized on a per-collection or per-division basis, strongly attached to a specific presentation, stored in different formats at different locations, and sometimes in multiple iterations. In some of the very earliest collections, the only identifiable source of descriptive data was in the record set on the web server that was feeding the presentation. Changes, updates or re-exports of the data required institutional knowledge or step-by-step reconstruction or transformation, but the project teams who had created the contents had largely moved on to other activities and did not always recall the decisions made or actions taken. The valuable digital objects that took so much time, effort, and resources to create were at risk of becoming unusable in the long term. In addition to that, the highly customized, inconsistent and handcrafted data were not particularly flexible or shareable. They could not be mapped to new standards such as MODS, or made accessible through tools such as OAI. This prevented it from being used easily or practically with any other presentations, or perhaps even within AM in the event that AM changed its structure.

Salvaging the Data

Staff members who developed the AM cookbook and the AMnonmarc storage record and accompanying tool were also the group largely responsible for locating or correcting the legacy record sets. Through their work transforming data sets and standardizing and normalizing the AM production processes, they developed expertise in the descriptive data and in the digital content data as they functioned within the application. They also provided production assistance to new non-AM digital library projects intending to include digital objects created during AM production. This exposure to new efforts helped keep them current with new standards or changes in the field. They began to recognize the potential hazard of leaving the legacy data as they had been created and stored, where they took great effort to update, were not supported, and could not be leveraged for re-use elsewhere. Likewise, any future development of the AM application could be restricted by the exceptions and customizations required for the legacy data to function.

The staff realized that with dedicated time, effort and resources, it might be possible to bring the legacy data into conformance with both the AMnonmarc storage record and the cookbook models. This would create a normalized set of non-MARC digital objects that could be stored in a supported tool, used with the existing AM application, and transformed into multiple standard formats. In 2004 a team began to work on upgrading the data, but the startup phase was slow. The major development during this period was the translation of the AMnonmarc storage format from its original RDBMS data model form into an XML schema that also incorporated METS-like structmaps with pointers to the content files. Once that was accomplished all new AM collections using AMnonmarc, though still created and stored in the database tool, would be delivered to the programmers as AMnonmarc XML. To increase flexibility, all AMnonmarc records were also required to have machine-readable dates, handles (persistent URLs), format terms, and web collection identifiers. These elements were added with the intent of increasing versatility for browsing, searching, and object re-use in potential future applications or upgrades to the AM presentation.

In 2006 the project applied the DLF Aquifer Implementation Guidelines for Shareable MODS Records[2] to the AMnonmarc data to increase its usefulness and shareability. This led to the addition of fields such as digital origin and genre terms to the AMnonmarc storage record field set. Unable to identify an existing set of controlled genre terms that included the diversity of genres within the AM collections, the project developed the Basic Genre Terms for Cultural Heritage Materials (BGTCHM)[3] and applied them to all the content being normalized under the legacy data project.

The upgrade is still underway, but the goals of increased sustainability, flexibility and shareability have already been met on a small scale. As part of the upgrade process, repeatable and automated processes were established to transform old record sets into AMnonmarc XML format. This will increase the sustainability of the objects by eliminating the need to locate, remember, or reconstruct past practices and data sets. The digital objects have the potential increased flexibility within the presentation programming from the addition of subjects, machine dates, and genre terms that are

[2] DLF Aquifer Implementation Guidelines for Shareable MODS Records http://www.diglib.org/aquifer/dlfmodsimplementationguidelines_finalnov2006.pdf

[3] Basic Genre Terms for Cultural Heritage Materials http://memory.loc.gov/ammem/techdocs/genre.html

available for future indexing. The objects are also transformable into OAI-MODS, OAI-DC, MODS, MODS-Aquifer, and METS, and other groups within the Library have been able to use subsets of these AMnonmarc XML data objects for new projects as well. One Ameritech partner requested a MODS version of the non-MARC records it had created for use in the AM presentation, and the Library was able to provide it easily because the record set had been upgraded to AMnonmarc XML as part of the legacy project. This was good validation to the legacy project team that the upgrades have succeeded in making the data more shareable as well as sustainable and flexible.

Continuing Challenges

The AM legacy data normalization project was designed to ensure that valuable digital content would continue to be used and made available to the public in the future. However, the work of salvaging data can be painstaking and tedious. In the case of the AM legacy data it is also largely invisible. Data underlying the presentation are being enriched, but the the presentation itself has remained consistent. Many users assume by the persistence of the look and feel of the presentation that the level of functionality created by data upgrade has always been there. Future upgrades to the AM application on the whole are out of scope for the legacy data project. However any future changes would be difficult or impossible were it not for the effort to salvage the seemingly unsalvageable legacy data.

Conclusions

Lessons learned from the early production of AM continue to be of service to digital production work today. Despite the creation of new standards and tools to address digital library production and display, the source content—whether digital or physical—and its associated description often exist in non-standard forms that must be manipulated or transformed before they can be integrated with new presentations, ingested into tools, or expressed according to standards. The Library of Congress received its first donated digital archive in 2003, and analysis of the contents indicated that born-digital content is also subject to idiosyncracies of organization, description and exchange (Library of Congress, 2005)[4]. In future production of digital library content, the difference may come not from having higher-quality source data, but from knowing at the outset what to expect and how to proceed based on past experience.

Acknowledgements

I would like to thank the AM data group, programmers and others at the Library of Congress for their support and hard work saving the data. Thanks also to Martha Anderson for the five rules of “Everything we needed to know we learned from digitizing” that provide such an irresistible framework for the sharing of lessons learned.

References

Libray of Congress. (2005). *Archive ingest and handling test (AIHT) final report*. 2005, June. Library of Congress. Retrieved November 28, 2008, from http://www.digitalpreservation.gov/partners/aiht/high/ndiipp_aiht_final_report.pdf

[4] See also (Madden, 2007) for more information about analysis of the content itself.

Madden, L. (2007). Digital curation at the Library of Congress: Lessons learned from American Memory and the archive ingest and handling test. *DigCCurr2007, April 18-20, 2007, Chapel Hill, NC*: Retrieved November 2, 2008, from http://ils.unc.edu/digccurr2007/papers/madden_paper_6-2.pdf