Digital Curation at the Library of Congress: Lessons Learned from American Memory and the Archive Ingest and Handling Test

Liz Madden, Library of Congress, OSI/IM/DCS, 101 Independence Ave SW, Washington DC 20540, emad@loc.gov

Abstract

The Library of Congress (LC) created the American Memory (AM) website in 1995 to provide public access to digitized versions of materials from the LC collections. The technical architecture, programming, and bulk of AM content were created before the development or widespread application of digital library metadata and data standards. The staff that produced much of this digital content did so without tools designed specifically for this type of work and with no existing model of a web-based digital library for guidance. Twelve years later the data and metadata driving AM are managed in the same framework in which they were created, and the effects of digital-curation decisions made early on are coming to light in efforts to sustain, use, and share this valuable body of content. The retrospective view provides an opportunity to identify approaches to data that contributed to the long-term sustainability of the contents, and approaches that were detrimental.

In 2003, LC staff applied many of these lessons learned to the analysis of a donated digital archive. The archive comprised diverse formats and metadata collected on the web from different donors. The LC analysis provided further confirmation of good and bad practices with regard to the maintenance of digital materials, and helped refine the list of useful skills and practices for digital curation staff “in the trenches.” This paper discusses the findings and conclusions of analyses of AM and the donated digital archive as they pertain to digital curation. It also suggests opportunities for digital curation curriculum development.

Introduction

The American Memory presentation (AM) was the product of the Library of Congress’s (LC) National Digital Library Program (NDLP), which was announced in 1994, the same year that Netscape Navigator 1.0 was released.[1] The first 4300 items went live in 1995, just in time for traditional dial-up providers AOL, Compuserve and Prodigy to begin to offer internet services.[2] In its first five years AM produced nearly 700,000 described items, stored approximately five million content files, and made nearly 90 web collections available to the public[3]. The staff involved in creating this digital resource came from diverse backgrounds; some had traditional library backgrounds and MLS degrees, others were specialists in content areas such as music, folklore or photography. There were programmers and people with knowledge ranging from HTML, to common desktop spreadsheet and database software, to general information management. Everyone had the common goal to create a digital library presence for the LC, but no guidelines, standards, infrastructure, or processes existed to provide them with a roadmap. They largely made it up as they went along, looking for wisdom in the experiences that all brought to the work from their respective backgrounds, paying close attention to new problems as they arose, identifying solutions, and then looking for the patterns in what succeeded and what did not.

By the year 2000, the World Wide Web had taken hold in society. Recognizing this, the LC made a commitment to continue the historical conversion work begun with AM and to expand into broader areas such as born-digital content and web capture activities. To date there are 35 TB of converted historical materials available at LC, representing more than 1.5 million described items.

The formation of the National Digital Information Infrastructure and Preservation Program (NDIIPP) highlighted the need to ensure that digital data could be created, captured, received, and maintained long into the future. The Archive Ingest and Handling Test (AIHT),[4] one of the first NDIIPP projects, provided an opportunity to apply knowledge gained from historical conversion work to a heterogeneous collection of data. The test data set was the George Mason University (GMU) 9/11 Digital Archive (DA), which comprised content donated to the archive via a website in the period after the attacks of September 11, 2001. GMU donated the archive to the LC in 2003 with the hope that LC could sustain it for the future. While analysis of the GMU 9/11 DA content as intellectual material was out of scope for the AIHT, the LC AIHT team recognized that the donation provided a practical, real-world example of what digital donations might look like in the future. With that in mind, the LC AIHT team performed its own separate analysis of the contents, which uncovered interesting implications of being the recipient institution for a digital archive composed of diverse formats and metadata donated by diverse "creators."

This paper focuses on the combined experiences of AM production and history, and the LC AIHT GMU 9/11 DA analysis. It approaches digital curation at a micro level—i.e., data workers and decisions involved in production work itself—rather than at the macro or institutional level. These data workers, referred to throughout this paper as "digital curators," routinely engage in such activities as data manipulation (wrangling), database and other tool development, technical assessments, specifications, automation of processes, and policymaking. Challenges for digital curators working with AM and the AIHT included permissions and rights, metadata creation, metrics and accounting, production tool development, problems of legacy data, and methods to ensure that data remained consistent, flexible, sustainable and shareable. The lessons learned from wrestling with all of these areas are as follows.

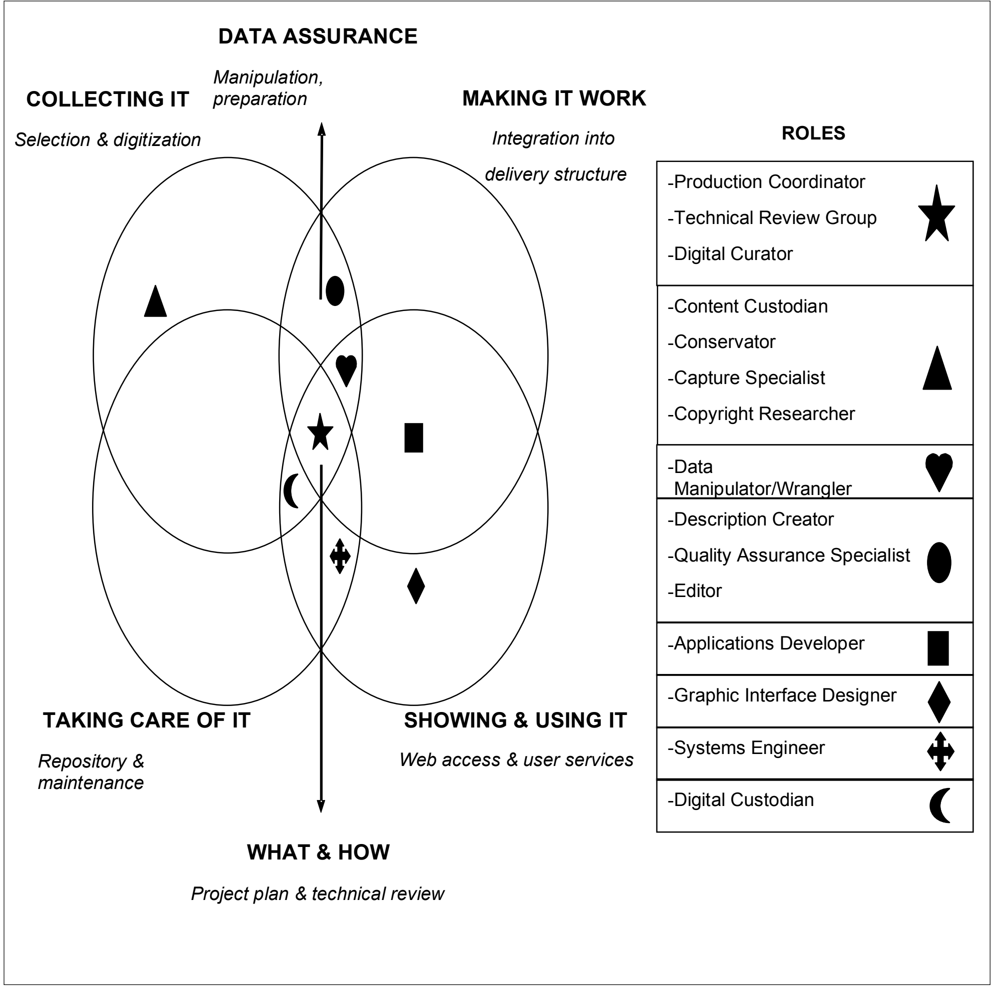
Lesson One: Digital Curation is Interdisciplinary

A particularly noteworthy lesson learned from both AM and the analysis of the GMU 9/11 DA is that digital curation by nature is interdisciplinary. The transformation or transfer of digital content for access and maintenance requires digital curators to have a basic understanding of a broad range of fields of expertise and to act as translators among the various experts in each of those areas. Digital curators are bridge builders who are able to communicate user needs to programmers and engineers, and engineering/programming constraints to users. They can come from diverse backgrounds because their value lies in their ability to understand digital content as both intellectual material and as data. They are parsers of data that travels from content creator to application and storage servers; they help develop requirements for production tools and systems; and they participate in the identification and establishment of practices and policies to ensure that data retains necessary intellectual and technical characteristics across all life-cycle stages. Above all, digital curators identify emerging patterns in the overall work; create strategies and methods to mitigate the negative outcomes; and work with others to create processes and infrastructure that lead to the most positive results.

In light of this, one of the biggest risks for digital curators is that their cross-disciplinary breadth may sometimes be mistaken for an absence of depth in any specific area of expertise. It may also provide a challenge to resource allocators and program planners who want to know where in the organization to hire staff to fulfill these roles. Experience in AM suggests that where these staff are located is less important to the overall success of digital curation of materials than the ability of the staff to communicate effectively across a wide audience. In reality, these digital curators often develop expertise in an in-between area such as "how to" instructions or simple tools or processes that become vital to the work at hand but would not be considered sophisticated by experts on either side. A single person is unlikely to be able to fulfill all digital curator functions; a coordinated team of individuals whose combined knowledge is broad and who can cultivate expertise in complementary areas will fare best.

Figure 1 below illustrates the cross-disciplinary nature of digital curation and digital curators. This interpretation of digitization work was developed out of the experience with AM. Each circle represents a major area of historical conversion and presentation work, and each intersection represents an area that requires communication across specializations. The more intersects, the broader the knowledge required to facilitate the work.

Figure 1. Aspects of Digital Collection Creation and Maintenance from American Memory[5].


DATA ASSURANCE
Manipulation, preparation
COLLECTING IT
Selection & digitization
MAKING IT WORK
Integration into delivery structure
TAKING CARE OF IT
Repository & maintenance
SHOWING & USING IT
Web access & user services
WHAT & HOW
Project plan & technical review
ROLES
-Production Coordinator -Technical Review Group -Digital Curator ★
-Content Custodian -Conservator -Capture Specialist -Copyright Researcher ▲
-Data Manipulator/Wrangler ♥
-Description Creator -Quality Assurance Specialist -Editor ⬮
-Applications Developer ▮
-Graphic Interface Designer ◆
-Systems Engineer ✥
-Digital Custodian ☾

Lesson Two: "Everything we needed to know we learned from Digitizing"[6]
Several major themes emerged during AM production that also held true in the assessment of the GMU 9/11 DA. Regardless of whether the goal is historical conversion or receipt of born-digital content, these five "commandments" of digital work have been borne out repeatedly through much trial and error and well-intentioned attempts to circumvent them.

Know thy Data
Understanding all aspects of the data and the expectation for its use and preservation is crucial to all stages of the digital life cycle. Appreciation for the intellectual or cultural value of material helps to inform the decisions about acceptance of donated content, digitization specifications for

converted materials, presentation or description requirements, and schedules for backup or integrity checking. Knowing how it will be used and stored in systems affects how it gets transformed or manipulated during production or presentation stages such as descriptive record creation, data delivery, full text handling, indexing, display, or storage. Expectations for user interactions may shape decisions about venues for presentation and access, data sharing, and choice of standards to apply. Knowing what files and metadata constitute an "item" guides permissions issues, rendering, description, and maintenance. Weighing the intellectual or cultural value against technical limitations or costs of receiving, sustaining, or working with data in a certain format can help inform decisions about what projects to embark on and what content to accept—as in donated digital content—and to manage the expectation of services that can be realistically applied to it.[7]

AM and the AIHT showed how knowing the data can also help with reporting required for funding, program development or public interest. Managers often ask the questions "how many?" and "how big?" in order to predict or measure cost efficiencies, productivity level, or resource needs. The answer of how many or how large depends on the nature of the request and the person asking it. Researchers want to know how many discrete intellectual items are available for their use; server storage engineers want to know how many files and what scale of data they need to accommodate; contract managers arranging for digitization of materials want to know how many scans they have to pay for; and recipients of donated materials want to know how many files are being transferred. Program directors are interested in all of the numbers as well as an explanation of what each number means. Knowledge of the data feeds directly into knowing what number to provide to what audience.

Automation Means Letting Machines Do the Work

Automation is repeatable, consistent, and does not rely on human or institutional memory. Ideally in an automated system, a user can press a button and walk away know that the process will happen, repeatedly if desired, and what the outcome or structure of the output will be. The likelihood that a mistake in an automated system will be applied consistently across the board, rather than sporadically or unpredictably, is greater than with manual processes. Therefore when a mistake is identified, one can assume that it has a widespread occurrence and can apply the fix broadly. In an automated environment, human intervention can be directed at exceptions and not at routine processes. Having fewer variables means fewer places to have to do troubleshooting. In addition to this, automation relies on data input that is consistent and recognizable to the system. This requirement is a catalyst for the development of standards, and standards in turn give data creators guidance for the structure of their data. This cycle fuels necessary research and development in areas such as data transfer, metadata structures and elements, and repository development.

Despite the availability of standards, experience with AM revealed that many data irregularities happen at the "pre-standard" production stage, during idiosyncratic transformations of data from one format to another, or during the transfer of data between one system or tool and another. Errors occur before data is in a standard form for ingest, sharing, or presentation in an application or system, such as during the earliest stages of a historical conversion project, when the only existing description of the items is a word processing document that must be transformed into a normalized format for import into a database or transformation into a markup form. Because the data in its existing form is not suitable for the tools and processes that exist, the temptation to intervene manually for the sake of expediting the production work is often irresistible. The intervention itself may appear negligible or harmless because it is perceived to be a one-off activity. Small interventions such as an extra line of code added to a script, or manual searching and replacing in a word-processing document can add up over time. They also often repeat themselves, but because they were initially considered exceptional, one-off situations, there is no automated process or documentation in place to guide the handling of them.

The lesson learned in AM from the recurrence of supposedly unique situations and one-time events was to automate whatever could be automated and to document each step. Even automation as simple as creating database export macros or scripts using regular expressions to do searching and replacing increases the sustainability of the data. For some AM data these routines had to be created retrospectively to ensure that they could be run again. In other cases the AM data had to be changed and new processes written because of an inability to re-trace or repeat the original handling.

Exceptions to Rules Raise Resource Usage

At the beginning of AM production, teams of staff were assigned to specific divisions at LC to work on the digitization of collections for the website. All data filtered down to a single team of three programmers with a responsibility to make the data work in AM. With no prior experience in how to create digital objects from physical objects, each team did the best they could with the content they had. A couple of years into the project, certain patterns began to emerge. Teams with homogeneous physical collections developed standard data sets and production routines earlier than teams with heterogeneous content. Each new homogeneous collection reinforced lessons learned from the previous one, but each new heterogeneous collection introduced variables not yet seen by that specific team. Even when they had encountered a certain content type before, teams with heterogeneous collection types sometimes developed a different way to handle it because the intellectual context of that specific collection was different. Each different approach required custom handling by the programmers.

Customization reduces sustainability of data and scalability of production by requiring extra or special processes to interact with the data. The exceptions may occur at any point between data creation and data storage. They come in the form of extra descriptive fields that must be accommodated for a certain body of content, or in lines of code created to treat objects differently from other similar objects, or in various other forms. Each exception raises the level of effort for data analysts, data wranglers, and programmers who have to create new routines to handle the special circumstances. These exceptions often result in the creation of “legacy data.” This term applies to data that lags behind current or developing models and technologies. Characteristics of AM legacy data are that it is inconsistent across collections, highly customized, created on a per-collection basis, strongly attached to a specific presentation of a collection, and stored in different formats and in different locations. Changes, updates or corrections to legacy data require institutional knowledge or research, which decreases the sustainability, flexibility and shareability of the data over the long term. Legacy data is resistant to change and often cannot be repurposed easily or used with new systems or architectures.

The proliferation of custom code in early AM resulted in the creation of an AM production team and technical review group. The review group assessed project proposals to gather the resource needs involved in doing the project. The production team surveyed the practices and data of various teams and identified patterns among them in order to enforce more consistent practices across the board. The production team acted as a liaison between the programmers and the digitization teams and developed AM local standards to streamline the production work. These local standards were expressed in the AM Cookbook of Collection Assembly, which provided teams with a set of models and data specifications for object types in AM, and the AM nonmarc descriptive record. This record format provided teams with a consistent set of fields to use in the creation of descriptive records that were not MARC-based (non-MARC). In recent years this AM nonmarc record has been expanded, expressed in XML with mappings to standards such as Dublin Core, MODS and METS, and forms the basis for efforts to migrate legacy non-MARC data into modern structures to increase its longevity.

Exceptions encountered during the AIHT also illustrate the cost of special handling. Two such exceptions involved problems with file naming and MIME types--areas that are generally assumed to be straightforward and standardized. Differences in file-naming rules across platforms caused problems with the initial transfer of data from GMU to the LC,[8] and files were not copied correctly from one location to the next as a result. The LC AIHT team spent time analyzing what went wrong and re-working the method of transfer to accommodate for the differences.[9] Similarly, files did not validate according to their apparent MIME types, requiring additional staff time to examine the validation routines and the files themselves to identify the source of the discrepancies. These are two examples from a relatively small archive, but extrapolated out to larger bodies of content they could become significant.[10]

Interoperability Requires Compromise
Interoperability requires data to be consistent, to contain required elements, and to be structured in an expected way. As with customization, exceptions to requirements raise resource usage when data created according to one set of rules is moved into a venue that applies a different set of rules or applies the same rules differently. Even when a common transfer structure is used, data may still not be readily exchangeable. During the Phase Two of the AIHT, participants exported a marked-up version of the GMU 9/11 DA from their own systems. Despite the fact that three of the four partners chose METS as the export format, each partner had its own approach to METS. One partner suggested that “Clearly there would be some advantage to working toward at least some common elements for these processes. Establishing this type of agreement early in the project likely would have improved the efficiency and success of the export/import component of AIHT.”[11]

Early agreements about data are integral to collaborative digitization projects as well. Approximately 25 of the collections in AM contain content from collaborators outside the LC. Work with these partners highlighted the need to establish early on the delivery schedule and number of redeliveries or updates, the need for persistence of data structure for corrections or future updates, the designation of responsibility for “master” data, the presence of a minimal set of metadata elements, and an understanding of the rights or permissions for use of the contents by each partner. Such agreements were not in place for these collaborative collections, and as a result many resources have gone into re-doing existing automated processes for updated content that no longer conforms to its original delivery structure. These collaborative projects also struggled with issues of data synchronization that occurred when the intellectual portion of partner data—i.e., fields in the descriptive records—was enhanced or corrected to work more cohesively within the AM structure. Enhancements or corrections that do not exist in the content maintained by the data originator will not be present in any update to the same data, and the work is at risk of having to be redone at the point of redelivery. Agreements at the outset of a project as to who is responsible for maintaining the “master” record for the content can mitigate this risk, or at least allow for additional data-massaging resources to be included in planning for the project.

Diversity must be Recognized
A fully interoperable and homogeneous digital world is unlikely and in some ways undesirable. Heterogeneity of systems and platforms may help mitigate the risk of loss due to platform-specific hazards, obsolescence, pernicious code, etc., and for this reason is beneficial to the long-term survival of digital content. In addition to this, diverse communities have different

needs for their data and applications, and in order for them to create and sustain their content they must develop methods, standards and routines that work for them. Even users within the same community may have different needs, depending on what stage of the life cycle they work in. The key to handling diversity may be in recognizing when to accommodate distinctions and when to conform to existing structures. In some situations forcing the data into a standard form may decrease the overall sustainability of the data and processes. Likewise, on occasions when the content itself appears to be like other existing content and the inclination is to conform, the trick to handling the data appropriately may be in looking at it with new eyes.

AM collaborative projects introduced diversity issues arising from tool and infrastructure development at the partner institutions. When the collaborations were first begun, some of the partner institutions chose to work with AM because it provided a service and expertise that was not yet developed in the institution itself. The partners provided data in whatever form they had available to them—spreadsheets, databases, delimited text files. In some cases a partner provided data in multiple forms simultaneously because it maintained content in different internal systems. AM created transformation processes based on the form of the original deliveries. As digital library expertise grew and developed, these partners began to create a digital library infrastructure of their own. This infrastructure incorporated new content management tools such as CONTENTdm® into their environments. Some of the tools automated the export of data from the systems, but only in designated formats—MARC or Dublin Core, for example. Subsequent deliveries or redeliveries of data came in these new standard forms instead of the original format, and the transformation routines had to be completely redone as if it were a new collection. In cases where the new export form lacked the discrete metadata elements required by AM programming, the new data could not even be transformed. The partner had to go back and figure out how to re-create the elements in the original data set out of the new tool being used because the constraints of the new tool were too limiting for the requirements of the project. In another case a union catalog tool used by a partner did not have de-duping capability and duplicate records were included in the delivery to LC, where the actual de-duping took place. This problem of having only controlled access to backend data was a consistent theme throughout AM production, both internal and collaborative. Some of these transfer issues might have been avoided if the data had been accessible for manipulation and export at a more granular level and not just as part of a packaging routine.

Experience in the AIHT illustrated how new content types require different approaches, and how applying practices associated with one content type to a new type of content can influence the interpretation of the data. The GMU 9/11 DA delivery included the data files themselves, documentation from the GMU team, a transfer manifest requested by the LC, and a MS Access database containing available metadata about the archive contents. A checksum comparison of the files revealed that 5% of the objects were duplicate and that these 5% were distributed across 146 different collections within the archive. In the context of historical conversion, this might have led to de-duping activities. However, the documentation that accompanied the GMU 9/11 DA delivery suggested that with web-collected content, duplicate checksums could arrive from multiple donation of the same materials from different contexts. This meant that checksum alone was not sufficient for identifying independent objects.[12]

The idea of content interrelationship and dependency also came to light during the LC AIHT analysis. Most of the archive contents were web-based, which meant that contents could be accessed via a link on a web page if the linked material was also part of the archive. Understanding and maintaining these internal relationships is an important part of preserving the ability to access an archive's materials in the future. Of the archive content, 29% was HTML, and 8% of the HTML contained HREFs that pointed outside the archive. Again, this illustrates the importance of context and interrelationships for digital materials. If the pointers go to additional content in the form of images, sound files, video, text, etc., then there can be no guarantee that content will be available in the future to help frame the materials within the archive. The same is true if the pointers go to scripts that create more content. Treating this sort of content the same as converted content risks losing the context and dependencies from which the data derives much of its intellectual value. In this case the diversity must be accommodated.

Lesson Three: Digital curators are an excellent investment
If policies, systems, and standards create a blueprint for a trusted digital repository and the digital content constitutes bricks, then digital curators are the mortar between the bricks. In order to be effective they must get mixed up in the work itself. Exposure to and participation in a broad range of activities is at the heart of their work. Digital curators are present at most discussions and meetings about data during each stage of the life cycle. Their role is twofold: first, to assess and plan for the transfer of data from one stage of its development to the next, and second, to represent expert stakeholders who are not present in a given meeting but whose needs or constraints have an effect on the data at another stage. Because of their involvement throughout the entire process, digital curators also often end up holding valuable institutional, programmatic or field-wide knowledge. They can predict future problems based on past experience, and they can recognize trends and know when to intervene to prevent data loss. They can inform requirements for tool development or new policies. In standards or systems planning exercises, digital curators are able to report on what elements are most likely to be present and which ones will be more difficult to require. They are excellent troubleshooters because they have seen the workarounds that users will do in the absence of a suitable tool or process for their work.

Digital curators who know their data and understand their organizational structure and environment are in a good position to identify at-risk or unsustainable content, or candidates for upgrade or transformation. They can identify opportunities for compromise or help assess the true interoperability of purported interoperable systems because they understand how their data relates to all aspects of the digital life cycle. They can say whether necessary elements are present or absent, if they are in or can be manipulated into the expected form, and if the processes to do so can be automated. They may be useful consultants for agreements and specifications for data transfer or data sharing projects. Exposure to and involvement in the full spectrum of an organization’s digital work is key to developing digital curators as a resource, but aspiring digital curators can cultivate certain skills outside the context of a digital library as well. A basic understanding of relational database concepts, data normalization, data modeling, simple scripting or programming, and systems analysis and design are all valuable skills for digital curation work. The following exercises, inspired by lessons learned from work in AM and the AIHT, might also help prepare digital curators-in-training for working in the trenches of a digital library.

- Hard Drive Processing
    - Find a hard drive that mimics a “normal user” hard drive with heterogeneous materials. Transfer, process, analyze, and create descriptive records for the objects on it
- Create descriptive metadata records from a non-EAD finding aid, word processing document, inventory database, spreadsheet, or any other list of contents for digitization. For extra credit:
    - Automate the transformation using simple scripts or macros of some kind
    - Create a desktop database or other system to store the data

    - Exchange transformed data with another student and ingest it into your system
- Distribute files created on old media (e.g., floppy disks), in software that is ten years old or older, or that is in a different character set and see who can read and identify them

Acknowledgments

This report is built on the work of all the digital projects and American Memory staff at LC, without whom no lessons would have been learned. Special thanks to fellow data wranglers and digital custodians Christa Maher and Timberly Wuester for their valuable contribution to these lessons learned, to Justin Littman and Dave Hafken, who provided technical expertise for the LC AIHT analysis, and to Michael Stelmach and Beth Dulabahn for reviewing this paper and providing helpful comments.

[1] Wikipedia. http://en.wikipedia.org/wiki/Timeline_of_computing_1990-forward

[2] Robert H Zakon. *Hobbes' Internet Timeline*, version 8.2. http://www.zakon.org/robert/internet/timeline/

[3] *Annual Report of the Librarian of Congress for Fiscal Year Ending September 30, 2000*. Library of Congress http://www.loc.gov/about/annrept/annreports/fy2000/loc_2000_pages_118-122.pdf

[4] For information about the AIHT, see http://www.digitalpreservation.gov/library/technical.html

[5] LC Technical Design Review Group, December 2001. Rev. March 2007

[6] Thanks to Martha Anderson, LC, for "Everything we needed to know we learned from digitizing"

[7] See Richard Anderson et al., "The AIHT at Stanford: Automated Preservation Assessment of Heterogeneous Digital Collections." DLib Magazine, Dec 2005. Volume 11 Number 12. http://www.dlib.org/dlib/december05/johnson/12johnson.html for one AIHT partner's perspective on preservation expectations for digital content.

[8] Clay Shirky, "Library of Congress Archive Ingest & Handling Test Final Report" June 2005 http://www.digitalpreservation.gov/library/pdf/ndiipp_aiht_final_report.pdf, pp. 13-14

[9] Clay Shirky, "Library of Congress Archive Ingest & Handling Test Final Report" June 2005 http://www.digitalpreservation.gov/library/pdf/ndiipp_aiht_final_report.pdf, pp. 19-20

[10] Clay Shirky, "AIHT: Conceptual Issues from Practical Tests" D-Lib Magazine, December 2005. Volume 11, Number 12. http://www.dlib.org/dlib/december05/shirky/12shirky.html

[11] Tim DiLauro et al. "The Archive Ingest and Handling Test: The Johns Hopkins University Report" D-Lib Magazine, December 2005, Volume 11, Number 12. http://www.dlib.org/dlib/december05/choudhury/12choudhury.html

[12] From GMU project documentation: Marty Andolino and Jim Safley. *September 11 Digital Archive: Transfer of Archive from Center for History and New Media to the Library of Congress*. 8 Jan 2004.