



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

Liz Madden
Library of Congress
emad@loc.gov
DigCCurr2007
April 19, 2007

5/17/2007



What is a Digital Curator?

- A shepherd of data as it transitions from one stage of the digital life cycle to the next
- Someone who understands the dual nature of digital resources--as both intellectual resource and systems data
- A bridge or translator among experts in the various areas of digital library work
- A de facto resource for institutional knowledge about digital content creation, processes, and maintenance

Hot Topics for the Digital Curator

- Digital production workflow
- Data-in-the-raw: “pre-standards” stage
- Data transfer, transformation, manipulation
- Automation and repeatable processes
- Data flexibility, shareability, sustainability
- Requirements for production tools, storage systems and display applications
- Documentation and institutional memory

What do digital curators do?

- Analyze a body of digital content and make pragmatic and realistic assessments of services that can be provided for it
- Assist with digital production tools, processes and policies
- Ensure that data is structured to allow for appropriate access to or presentation of the content
- Ensure that data is structured to allow for appropriate production tracking, storage and maintenance of the content
- Identify places in the production workflow where digital content is at risk of becoming corrupt and propose solutions or mitigation strategies

Digital Curators are Interdisciplinary

There is depth in breadth

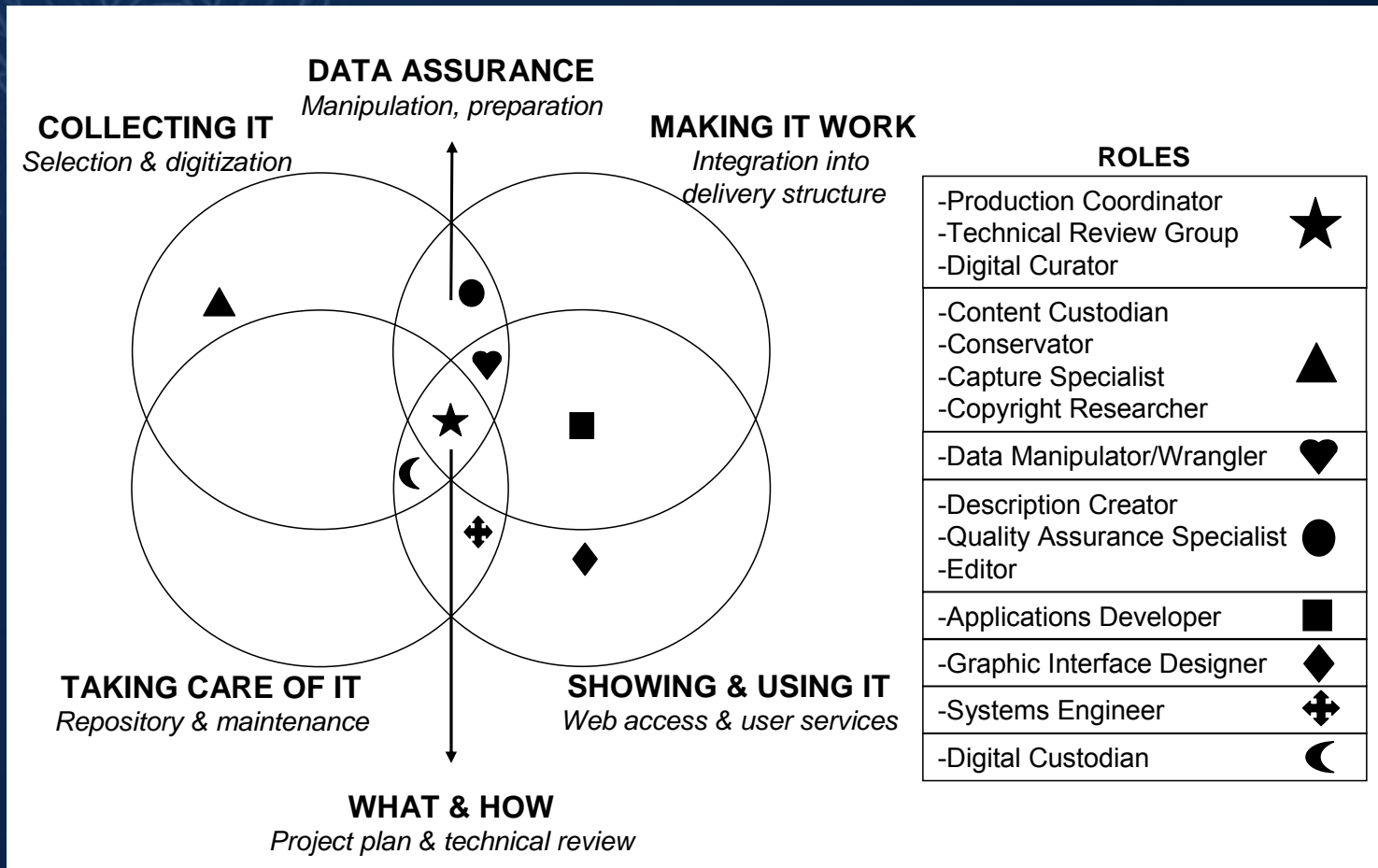


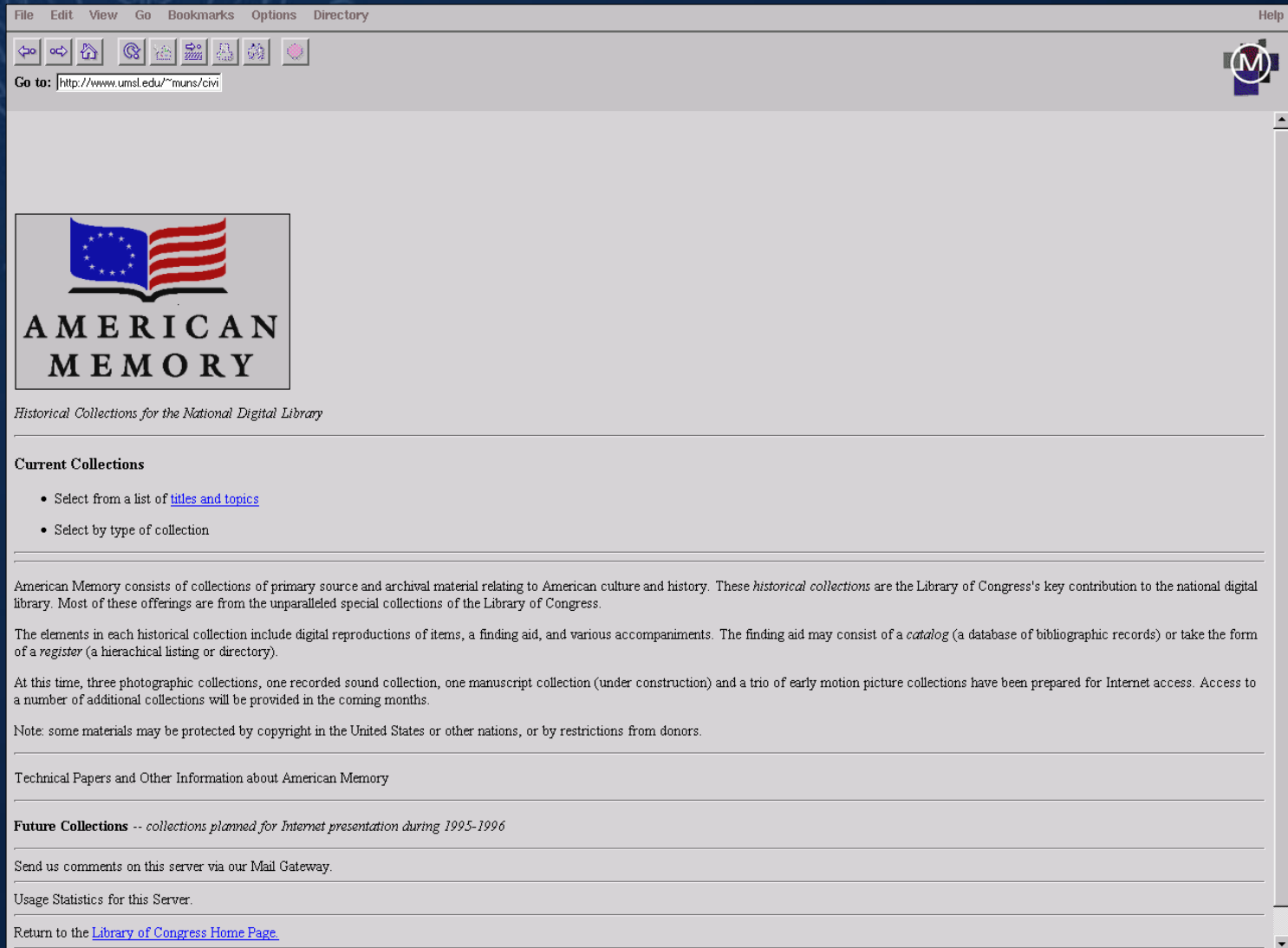
Figure: "Aspects of Digital Collection Creation and Maintenance"
Library of Congress, Technical Design Review Group, November 2001
(slightly revised, March 2007)

A Brief History of American Memory

<http://memory.loc.gov>

- Made its debut in 1995 with 4300 described digital items
- Produced approximately 700,000 described digital items in 90 digital collections between 1995 and 2000
- AM collections include content created by external institutions via the Ameritech Competition (begun in 1996)
- Currently contains more than 1.5 million described digital items and 35 TB of data
- Continues to grow


American Memory 1995



The screenshot shows a Netscape browser window with the address bar containing `http://www.umsl.edu/~muns/civi`. The page features the American Memory logo, which consists of a stylized American flag above the text "AMERICAN MEMORY". Below the logo, the text reads "Historical Collections for the National Digital Library". The page is organized into sections: "Current Collections" with a bulleted list of links for "titles and topics" and "type of collection"; a paragraph describing the historical collections; a paragraph about the elements of each collection; a paragraph about the current state of collections; a note about copyright; a section for "Technical Papers and Other Information about American Memory"; a section for "Future Collections" with a note about planned presentations; a link to a "Mail Gateway"; a link to "Usage Statistics for this Server"; and a link to the "Library of Congress Home Page".

File Edit View Go Bookmarks Options Directory Help

Go to: `http://www.umsl.edu/~muns/civi`



AMERICAN
MEMORY

Historical Collections for the National Digital Library

Current Collections

- Select from a list of [titles and topics](#)
- Select by type of collection

American Memory consists of collections of primary source and archival material relating to American culture and history. These *historical collections* are the Library of Congress's key contribution to the national digital library. Most of these offerings are from the unparalleled special collections of the Library of Congress.

The elements in each historical collection include digital reproductions of items, a finding aid, and various accompaniments. The finding aid may consist of a *catalog* (a database of bibliographic records) or take the form of a *register* (a hierarchical listing or directory).

At this time, three photographic collections, one recorded sound collection, one manuscript collection (under construction) and a trio of early motion picture collections have been prepared for Internet access. Access to a number of additional collections will be provided in the coming months.

Note: some materials may be protected by copyright in the United States or other nations, or by restrictions from donors.

Technical Papers and Other Information about American Memory

Future Collections -- *collections planned for Internet presentation during 1995-1996*

Send us comments on this server via our [Mail Gateway](#).

Usage Statistics for this Server.

Return to the [Library of Congress Home Page](#).

The Archive Ingest & Handling Test (AIHT)

<http://www.digitalpreservation.gov/library/technical.html#AIHT>

- Test the feasibility of transferring digital archives *in toto* from one institution to another
- Used George Mason University 9/11 Digital Archive (GMU 9/11 DA) as donated to LC
 - small, heterogeneous, real-world archive
 - 12+ GB
 - 57,000+ files
 - Created from content contributed via web interface
- Describe it, mark it up, ingest it, transform it, share it

LC as Recipient of GMU 9/11 DA

- The LC AIHT team performed an additional analysis of GMU 9/11 DA from the perspective of the recipient institution
 - What's in it?
 - What can we do with it?
 - How does it compare to other materials that we have experience with?

Everything we needed to know we learned from digitizing

- Know thy data
- Automation means letting machines do the work
- Exceptions to rules raise resource usage
- Interoperability requires compromise
- Diversity must be recognized

Know thy data

- Understanding the data as content helps identify
 - what the item is and what its boundaries are
 - how it might be best presented
 - uniqueness
- Understanding data as data helps ensure that it
 - contains the structure and elements to inform the presentation
 - Is in a structure that can be maintained as needed
- The same piece of data can function differently depending on the context
 - Title in scan list identifies a piece of physical material that must be scanned
 - Title in a production tool or descriptive record describes what the item is intellectually
 - Title in an application is a field whose data is used for
 - Display in a certain location
 - Indexing/searching/discovery
 - Creation of browse lists
 - Duplicate items in the GMU 9/11 DA differ by context
- Know how a change to a metadata field's value in order to accommodate one context can affect the use of that field's data in another context

Automation means letting machines do the work

- Machines are more regular than humans and more likely to make the same mistake the same way twice+
- Content in transition is at risk for errors
 - Transformation of existing description from one format to another
 - Manipulating a word processing file, spreadsheet, desktop db, etc., for loading into production tool or metadata schema
 - Transfer of content from one server architecture or platform to another
- There is often a strong temptation to do things manually for expediency
 - There's no such thing as "just this once"
 - No guarantee you'll remember what you did, and others may not know
- Automate whatever can be automated and save resources for exceptions
- Any automation counts
 - Database or word processing macros, simple scripts
- Document everything, especially the parts that cannot be automated

Exceptions to Rules Raise Resource Usage

- Customization is a kind of exception
 - AM collection model
- Customization reduces sustainability
 - May rely on institutional memory or documentation (often absent)
 - Solutions to identified problems cannot be applied broadly
- Exceptions compromise scalability
 - Look for patterns to identify true exceptions vs. existing models
 - E.g., “Parts of a whole” objects can come in a variety of forms
- “Standard” ≠ “Consistent” or “Interoperable”
 - Approaches to or interpretations of standards can vary on a local level to accommodate a particular environment
 - Standards that function in one environment may not function well in another
 - MIME type and file validation
 - File naming and environment differences

Interoperability requires compromise

Collaborative projects are challenging with regard to

- Consistency and persistence of delivery formats
- Data synchronization
- Updates
- Data transfer
- Metadata standard \neq data model
 - Use of a metadata standard does not guarantee interoperability with a system using the same standard
 - Standards are sometimes interpreted and/or applied in diverse (i.e., “nonstandard”) ways
 - Sometimes a data model needs to serve multiple standards
- Data that functions in one environment may not function the same way in another
- Up-front agreements about delivery structures, manifests, schedule, formats, and expectations for use and maintenance contribute to long-term sustainability of the data

Diversity must be recognized

- Heterogeneity is a fact of life
- There's no such thing as "one size fits all"
 - Requirements for data depend on the audience/user and the technical environment/application
 - Data created in one environment for one type of user may not contain all the elements needed for use in another environment for another type of user
 - Sometimes you can't declare that one environment is the only place data will be used
- Some things that look like they can be "homogenized" are the hardest to control or apply
 - Vocabularies (subjects, geography, genre)
 - Metadata fields
 - Local uses

Descriptive Metadata

- Descriptive metadata is important for resource discovery
- Many items of interest lack an existing descriptive metadata that can serve the presentation
- Existing descriptive metadata comes in all shapes and sizes
- Creation or enhancement of descriptive metadata is often done by non-catalogers
- There is overhead associated with leveraging existing descriptive information for digital production tools or presentations
 - Existing records may be in spreadsheets, word processing documents, inventory databases, non-relational databases, etc.
 - Data manipulation/wrangling for data in non-usable form
- Descriptive metadata must sometimes serve multiple environments that have different requirements
 - Vocabularies
 - Local fields or local use for existing fields that must be preserved

Exercises for aspiring digital curators

- Process a hard drive
 - Find a hard drive that mimics a “normal user”
 - Transfer, process, analyze, create descriptive records for its contents
- Create descriptive records from a non-EAD finding aid, word processing document, inventory database, spreadsheet, contents list, etc.
 - Automate the transformation
 - Create a solution for storing/using it
 - Do an exchange with someone else and see if you can ingest other data into your system
- See if you can identify files
 - Stored on old media
 - Created in old software
 - Based in a different character set

Other helpful skills for a digital curator

- Ability to run or create processes and tools to organize and/or manipulate data
 - E.g., Databases, XSL, Scripting, Regular expressions
- “Conversant” in
 - Relational database concepts, data normalization
 - Data modeling
 - Systems analysis and design concepts
- “Interdisciplinary” view of digital library work
 - What factors in one area have an impact on another area