Library of Congress Audio-Visual Prototyping Project
Carl Fleischhauer, National Digital Library Program
Outline of talk (omitting diagrams) for CNI Task Force Meeting, San Antonio, December 7-8, 2000

Background

National Audio-Visual Conservation Center
-- New Library of Congress facility, development led by the Motion Picture, Broadcasting, and Recorded Sound Division and National Digital Library Program
-- Support from the David and Lucile Packard Foundation and the Packard Humanities Institute
-- Will be in Culpeper, Virginia, 70 miles from Washington
-- Planned to go operational 2004
-- Development and prototyping during 2000-2003

Project Motives
-- Alternative preservation approach
-- Analog magnetic recordings: tape-to-tape copying never very good idea
-- Cessation of manufacture of tape and tape recorders
-- Risk of deterioration of tangible born-digital, e.g., audio CDs
-- Opportunity to begin investigation of intangible born-digital content, e.g., MP3
-- Provide access
    -- LC researchers on Capitol Hill
    -- Collections in Culpeper
    -- Possible future authorized research sites
    -- Limited outreach, most items protected by copyright

Project Themes
-- Content streams to preserve
    -- Reformatting analog and tangible digital - major element
    -- Processing intangible born digital - minor element
-- Preserving digital content from both streams

NOTE: Traditional preservation copying to continue until new approach is accepted

Digital Content

Reformatting Questions
-- Preservation quality bitstreams
    -- Bitstream type? PCM audio? Bit-mapped images? Video?!? Native formats for tangible born digital?
    -- File format? Industry "standard" WAVE and TIFF? CD-A to WAVE?
    -- Quality level? Samples per second? Pixels per inch? Bits per sample or pixel? Accept native quality level for born digital?

NOTE: I am happy to discuss in extreme detail this if the attendees are interested, or ask me later.

Intangible Born-Digital Questions
-- Secondary prototyping element today
-- Analyze
    -- Unpack? Reformat? Future system emulation?
    -- "Transform into persistent object" (NARA/UCSD)

Metadata and Object Structure

Metadata Categories

Bibliographic or intellectual metadata
    -- What content is this?
    -- Intellectual metadata supports discovery
-- Sample fields
    -- title, creator (author), publisher, disc label name and number, subjects, original physical desc
    -- table of contents to a long audio file

Administrative metadata
- What do I need to know to manage this object?
- Supports content preservation (e.g., archiving, migration, emulation)
- Supports control of access, plans call for on-site limit to copyrighted content

-- Sample fields for general administration
- access category and/or rights information
- about the source item that was reformatted
- persistent name (URN, handle)
- who/how digitized
- who is responsible for object management

-- Sample fields to support migration and more
- encryption, internet media type, file extension
- checksum (to verify file integrity)
- technical data re: bitstreams ("format metadata")
  - sample rate, bit depth, color space, compression, targets, spatial resolution, pixels horiz& vert, watermark, and more
- system emulation requirements (future)

Structural Metadata
- How does this object fit together?
- LC experience strongest re: reformatted content

-- Sample data
- Express hierarchy
  - primary, intermediate, and terminal levels
- Express relationships
  - "I am the high resolution version of page 3"

Reformatting Hierarchical Content
-- Illustration: pop music album
-- Two-sided 12-inch disc with printed labels
- 8 musical selections
- 4-page booklet (fake for our mockup!)
- Album cover art and text front and back
- Master and service reproductions

-- Total 32 or more digital files

[Diagram in slide show]

Metadata Input and Output: Goals
-- Emerged in process, difficult to synch
-- Database to capture metadata
- structured for efficient data entry

-- "Archival" XML document for long-term retention
- complete data set, "migratable" as needed
- compare OAIS archival info package

-- "Presentation" XML document
- streamlined for good fit to user interface
- compare OAIS dissemination info package
- Note: copyrighted content, local client, no WWW

Metadata Input and Output: the “Capture” Database
-- MS-Access Database
-- Circa 150 fields in a dozen tables
-- Working to automate data grab from bib records, file headers, and directory lists of files

Metadata Input and Output: the Archival Object
-- XML Concept from the Making of America 2 project at Cal Berkeley
-- Promising for archival object, contains complete set of metadata in migratable XML form
-- Current XML DTD needs to evolve/expand to embrace audio-visual elements

Metadata Input and Output: Presentation Object
-- We have a proof-of-concept interface client and dataset from special XML document
-- Desire: the special XML document should be subset of MOA2 document type, derivable from it

Metadata Input and Output: What is Difficult?
- Too darn many fields
  - 150 "possibles" in our set at this time
  - includes reformatting documentation
- Evolving fit with MOA2
  - add new elements for audio-visual content, cannot use DTD out of the box
- Cumbersome data creation and transformation
  - proof-of-concept mode is "handmade"
  - transformation may be seen as OAIS ingestion, downstream of production

Content Preservation

Today
- For now, working in a UNIX storage network world
- Masters in one set of filesystems, service copies in another
- Essence bitstreams and XML in online or nearline storage are the "preservation copies"
- Archived copies on offline media are "protection copies," remake periodically

Tomorrow
- AV project produced conceptual design for repository
- As this work proceeded, we studied two other models:
  - University of California, San Diego supercomputer center Persistent Archive Design
  - OAIS reference model (we now borrow its general concepts and terminology)

Under way
- defining ingestion and AIP(s)
- defining access and DIP(s)
- AV has special requirements

Deferred
- study of core OAIS elements archival storage and data management
- general LC enterprise development planned

Repository Design Gleanings
- System supports movement of content to other systems.
- Digital objects should be independent from the chosen storage manner or medium.
- Indirection provided by a "storage resource broker" (San Diego)
- Persistence service provides means to dynamically adapt new storage devices into the repository's infrastructure (AV concept design)
- Use metadata represented as XML documents which can be mapped to relational or object-oriented databases
- Allow for changes to metadata DTDs or schema evolution
- Provide for security, authentication, and authorization
- Members of federated archives need to manage both local and global user authentication and access management
  Repository Design Gleanings
- Incorporate "business rule engines" to permit changing workflow or other actions without extensive programming
- To overcome content obsolescence, consider
  - Modules to accomplish content transformation and migration (AV concept design)
  - Employing "persistent objects" that do not require migration
  - Emulating other or obsolete systems

Web Sites
- LC audio-visual prototyping project
  - http://lcweb.loc.gov/rr/mopic/avprot/avprhome.html
- LC enterprise-wide Digital Repository planning
  - Features metadata tables
  - http://lcweb.loc.gov/standards/metadata.html