Digital Audio Preservation at Indiana University (continued)

Jon Dunn
Digital Library Program

Digital Library Brown Bag Series
Indiana University
May 6, 2005
Digital Library Program
Interest in Sound Directions

- Experience with audio access (Variations/Variations2)
- Build knowledge in audio preservation
- Digital Library Repository Service project
  - Sound Directions serves as preservation pilot
IU Context: Technical Infrastructure

- Massive Data Storage Service (MDSS)
  - Central UITS resource
  - 1.6 petabytes, disk + automated tape
  - Mirrored between IUB and IUPUI

- Digital Library Repository Service
  - Under development by DLP
  - Access and long-term storage for digital library objects
  - To be based on Cornell/UVa Fedora software
Mass Storage Systems

- High-capacity, high-performance data storage
- Hardware
  - Servers
  - Automated tape libraries, e.g. IBM, Storagetek
  - Spinning disk
- Software
  - HSM: *hierarchical storage management*
  - IU uses HPSS (*High Performance Storage System*) from IBM
Mass Storage Systems

• Typical features
  • Bit-level storage and retrieval of files
  • Security: authentication, authorization
  • Mirroring of data between sites over a network
  • Migration of files to new media types
• Is that enough for digital preservation?
Data Persistence

- Key is *migration*
- Keeping the bits alive
  - Physical media
  - Logical media format
- Keeping the bits understandable
  - File format
  - Metadata
- Digital data must be *actively* managed
Digital Objects: More than just files

Example: Electronic Book

- Hi-res page image files (TIFF)
- Delivery page image files (JPEG)
- Text transcription (TEI/XML)
- Metadata

Hi-res page image files (TIFF)
Digital Objects: More than just files

Example: Sound Recording

- Delivery audio files (MP3 or other)
- Hi-res audio files (Broadcast WAVE)
- Images of labels, jacket, box, etc.
- Metadata
Digital Objects: More than just files

Example: Archival Collection
Digital Object Repository

- A system (hardware and software) in which to deposit digital objects (files and metadata) for purposes of access and/or long-term storage.
Digital Object Repository Purposes

• Access
  • Web access to digital files and metadata
  • Services/applications for searching, browsing, transformation, etc.

• Preservation
  • Secure storage for digital files and metadata
  • Services for file integrity checking (using checksums), migration, conversion, etc.

• Some repositories are single-purpose; some are dual-purpose
DL Object Repository

Users and Applications: Access and Management

Repository System

- Preservation version in mass storage system
- Delivery version(s) on web or streaming server
- Metadata records
OAIS: Open Archival Information System

• Conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term
  • “A common framework of terms and concepts … to provide long-term preservation of digital information”
  • Origins in space science community
• Discusses interactions that producers, consumers, and managers have with a repository
• Basis for much current thinking on repositories in digital library community
  • e.g. OCLC/RLG PREMIS project
OAIS Packages

- SIP: Submission Information Package
  - Ingestion into archive
- AIP: Archival Information Package
  - Internal storage in archive
- DIP: Dissemination Information Package
  - Delivery to consumer
    (could be another archive)
Components of an Audio Preservation Package

- Audio files: masters, derivatives
- Image, text files for accompanying material
- Metadata
- How to encode?
METS: Metadata Encoding and Transmission Standard

- METS, a Digital Library Federation initiative, attempts to provide an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users).
METS: Metadata Encoding and Transmission Standard

- “Glue” that holds a digital object together
- XML document which contains or references:
  - Descriptive metadata
  - Administrative metadata (including technical and digital provenance)
  - Structural metadata
  - Content files
- Can incorporate any metadata that can be represented as valid XML: e.g. MARCXML, MODS, Dublin Core, AES SC-03-06-A
Metadata

• Technical metadata
  • Records technical characteristics of digital object and its components

• Digital provenance (process history) metadata
  • Records actions taken on the digital object: digitization, conversion/transcoding, quality control checks, etc.

• AES task force SC-03-06-A working on METS-compatible schemas for both
Descriptive and Structural Metadata

- **Descriptive metadata**
  - Supports discovery (searching, browsing)
  - ATM traditional practice: collection-level MARC records
  - Preservation package: MARCXML?

- **Structural metadata**
  - Supports navigation
  - Information on content and location within audio files
  - ATM traditional practice: printed tape inventories
  - Preservation package: EAD or METS structMap+MODS or ???
Descriptive and Structural Metadata

- Not a clean dividing line between the two
- Different needs for access and preservation
- Granularity of preservation objects: collection or physical tape?
Sound Directions: Tools

- Harvard technical metadata and process history tools
  - Adapt to IU needs, including: port to Windows, handle WAV files
  - Release as open source
- Workflow database
- Preservation package validation, ingestion and dissemination tools
Digital Audio Archives
Project: DAAP

• Partnership with Digital Knowledge Center at Johns Hopkins University Libraries

• IU participants
  • Digital Library Program
  • Cook Music Library
  • School of Music Department of Recording Arts

• 2 year IMLS grant, began October 2004

• Explore efficient methods for high quality digitization of audio collections

• Focus on IU School of Music performance archive in Cook Music Library
IU School of Music Performance Archive

• Formats:
  • Transcription discs (1938-1956)
  • Open reel tape (vast majority)
  • DAT
  • CD-R

• Rich content
  • Student ensembles, faculty recitals, student recitals, guest artist performances

• 11,500 cataloged items in Music Library

• Project will digitize about 1800 hours
  • Selection criteria: technical, rights, cataloging, …
DAAP Rights Issues

• Two layers of copyright
  • Musical work
  • Recording
• Recording copyright held jointly by the entity making the recording and the performers
• No legal agreements with student performers
• What can be made accessible?
DAAP Plan of Work

• Establish new audio lab in Music Library
• Develop and carry out digitization workflow
  • Multiple simultaneous transfers
  • Quality control
• Enhance, create metadata
  • Additional access points
  • Technical/administrative metadata
• Load into Variations2 (and preservation repository)
• Assist JHU with evaluation of automated tools
Johns Hopkins Tool Development

• Automated quality control
  • Checking for typical problems, e.g. clipping
• Automated structural/descriptive metadata creation
  • Silence and applause detection
  • OCR on concert program
After DAAP

• Retrospective:
  • Conversion of remaining materials
  • Transcription discs

• Prospective:
  • Direct deposit of recordings into Variations2 and preservation repository
Questions?

• Contact information:
  • Mike Casey, micasey@indiana.edu
  • Jon Dunn, jwd@indiana.edu

• Sound Directions web site:
  • http://www.dlib.indiana.edu/projects/sounddirections/

• Presentation slides:
  • http://www.dlib.indiana.edu/workshops/bbspring2005.htm