Flipping the Switch: Lessons Learned from a Major Digital Library Migration Project

Jon Dunn, Ryan Scherle, and Mark Notess
Digital Library Program
Indiana University
DLF Fall Forum - November 7, 2005
Outline

- Background
  - Review of Variations and Variations2 digital music library systems
- Migration
- Post-migration experience
- Lessons learned
- Future directions
Variations

- Digital library of sound recordings and scores in IU Cook Music Library
- Online in April 1996
- 10,000 recordings; 200+ scores
- Access to course reserves, general, and special collections
- Available from workstations in Music Library and selected on-campus locations
- Jointly supported by Digital Library Program and Music Library
Variations

William and Gayle Cook Music Library

T351 - PERMANENT RESERVE
Reserve Recordings

Adams, John, 1947.
<Orchestra music. Selections>
The Chairman dances: foxtrot for orchestra; Christian zeal and activity;
Two fanfares for orchestra; Common tones in simple time <sound recording>/
CD A214 4e
Listen to online copy in VARIATIONS

Bartók, Bela.
<Duetos, violins>
44 Duos for Violins (1931) <sound recording>/ Bela Bartok -- Loosdrecht, Holland. Sound: N.Y.: Distributed by Qualiton Imports, <1980->
CD B292 C2S1 1-3
Listen to online copy in VARIATIONS

Bartók, Bela, 1881-1945.
<Quartets, strings>
The string quartets <sound recording>: complete / Bartok -- <Hungary>:
CD B292 C4S-13
Listen to online copy in VARIATIONS

Bartók, Bela, 1881-1945.
(Music for Strings, Percussion, and Celesta)

Listen to online copy in VARIATIONS
Variations

Search Results

“leonard bernstein and http and sound” search found 137 titles.

Mark items for Print/E-Mail/Save

#1

Songs and duets; Arias and barcarolles [sound recording]

Bernstein, Leonard, 1918-

Vocal music. Selections

1 sound disc : digital ; 4 3/4 in.

copy available at B-MUSIC. Click DISPLAY for details.

#2

On the waterfront [sound recording]. Chichester Psalms; On the town

Bernstein, Leonard, 1918-

Selections

1 sound disc : digital ; 4 3/4 in.

copy available at B-MUSIC. Click DISPLAY for details.

#3

Medea [sound recording]

Cherubini, Luigi, 1760-1842.

Medea. Italian

2 sound discs : digital, mono ; 4 3/4 in.

Copy currently not available. Click DISPLAY for details.
Variations

Author/Composer, etc.: Horne, Marilyn.

Title: I will breathe a mountain [sound recording] : songs of Barber, Bernstein, Bolcom.


Physical description: 1 sound disc : digital ; 4 3/4 in.

URL: http://www.music.indiana.edu/cgi-bin/var/access?CAR9723

Label no.: 09026.68771-2 RCA Victor Red Seal

Language: Sung in English.

Performer: Marilyn Horne, mezzo-soprano ; Martin Katz, voice (in Nachspiel) ; William Bolcom, piano (in Nachspiel) ; Tokyo String Quartet (in Dover Beach).


Notes: Original works and arrangements, principally for voice and piano, and for voice with string quartet acc.; Nachspiel for 2 humming voices and piano ; Compact disc.

Program notes by Steven Diller, biographical notes on performers, and texts in English (16 p. : ports.) inserted in container.

Contents: The daisies : op. 2 no. 1 (1:05) ; A nun takes the veil : op. 13 no. 1 (1:41) ; Bessie Bobtail : op. 2 no. 3 (3:14) ; The secrets of the old : op. 13 no. 2 (1:13) ; Sure on this shining night : op. 13 no. 3 (2:36) ; I hear an army : op. 10 no. 3 (2:26) ; Dover Beach : op. 3 (8:30) / Samuel Barber . . . I will breathe a mountain (24:21) / William Bolcom – My house (3:00) ; So pretty (2:21) ; Greeting (2:41) ; Take care of this house (1:55) ; Rabbit at top speed (1:19) ; Sonnet : What lips my lips have kissed (4:44) ; Nachspiel (2:04) ; Dream with me (3:31) / Leonard Bernstein.
Variations
Pre-Variations
Variations
Variations2 Project Overview

- R&D project funded by DLI-2 grant from NSF/NEH (2000-2006)
- Establish a digital music library testbed system supporting multiple formats: audio, video, score images, score notation
- Develop multiple interfaces for specific user applications in the music library and the classroom
- Conduct research in metadata, usability, copyright, and networking
Expanding on Variations

Variations2 expands on Variations by:

- expanding representations of music in other media
- creating additional metadata and new software tools for enhanced access, synchronization and navigation
- creating tools for pedagogical use
- expanding access locations
Variations2 Overview

Collection Metadata

Access Control, Bookmarks

Users
- faculty
- students
- librarians

Content
- digitized audio
- scanned scores
- encoded scores

User Interface

Variations2
Users can bookmark locations in tracks, send tracks to a personal playlist.
Score Viewer

Users can page through scores, bookmark pages, zoom in and out, and switch to a two page view.
Score Annotation

Users can also annotate scores and save or print the annotations. Annotations scale with page zooming level.
Any recording can be synchronized with any score of the same work for easy navigation and even automatic page turning.
Users can create visual timelines of any audio segment, annotating and labeling the diagram. Audio playback of the timeline displays the successive annotations. Diagram and annotations can be exported to a web page automatically.
Personalized Playlists

Users can create personal playlists from different albums.
Lesson Editor (experimental)

With Variations2 you can page through a Beethoven score, listen to Tito Puente, or diagram a sonata, all with content residing in the Variations2 digital library.
Users can search for music using music-related fields, such as performer, key or media format.
Data Model Example

**CONTRIBUTORS**
- Horowitz, pianist
- Uchida, pianist
- Mozart, composer
- Broder, editor

**WORKS**
- Sonata K. 279
- Fantasia K. 397

**INSTANTIATIONS**
- Sonata K. 279 recorded in 1965, Carnegie Hall
- Fantasia K. 397 recorded in 1991, Tokyo, Suntory Hall

**CONTAINERS**
- CD: Mozart, Piano Works
- Score: Mozart, Piano Fantasia K.397

Catalogers enter metadata for items, contributors, or works. Data can be imported from existing MARC records.
## Content and Metadata Comparison

<table>
<thead>
<tr>
<th></th>
<th>Variations</th>
<th>Variations2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Master files</strong></td>
<td>CD-quality WAV</td>
<td></td>
</tr>
<tr>
<td><strong>Delivery files</strong></td>
<td>384 Kbps MPEG-1 layer 2</td>
<td>192 Kbps MP3</td>
</tr>
<tr>
<td></td>
<td>(“MP2”)</td>
<td>28 Kbps AAC</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>MARC + “track file” stored</td>
<td>Complex metadata model in DB2</td>
</tr>
<tr>
<td></td>
<td>on filesystem</td>
<td></td>
</tr>
</tbody>
</table>
Delivery Comparison

- Variations
  - Automated tape library managed by Tivoli Storage Manager
  - Disk cache on IBM RS/6000
  - IBM VideoCharger streaming server
  - Custom Perl scripts on server, Player on client

- Variations2
  - Apple Xserve G5 with 7 TB RAID disk array
  - Apple QuickTime Streaming Server
  - Client-server Java application
Going Production

- Improve system reliability, stability
- Support out-of-library use
  - Student labs, other libraries, faculty offices, classrooms
  - Download for home use
  - Access control: authentication/authorization, integration with Student Information System (PeopleSoft)
- Develop support model
  - User guide, support processes
- Migrate content
- Migrate metadata
- Validate migration
- Get it done in less than 1 year
Planning for Migration

- Scope out and order new hardware
- Import all metadata into test database
- Design and build import process:
  - Do not require human intervention
  - Must be able to detect problems, and skip to the next item when a problem is detected
  - All output is logged, making it possible to diagnose/correct unanticipated errors
What we started with

- More than 18,000 .wav files in near-line “mass store” tape storage (HPSS)
- MARC records in our catalog
- Text “track files” with minimal bibliographic data and track listings:

Author=Beethoven, Ludwig van, 1770-1827.
Title=Symphony no. 5 in C minor, op. 67 ; Symphony no. 7 in A major, op. 92 ; Egmont
Copyright=(P) 1979 The Decca Record Company Limited

[Track 1]
WorkTitle=Beethoven, Symphony No.5 in C minor, op.67
TrackTitle=I Allegro con brio
Offset=0

[Track 2]
TrackTitle=II Andante con moto
Offset=432734
2 newly encoded QuickTime .mov files for each .wav file--192kbps and 28kbps of each "media object"

Database records for each new CD or Side, plus a "container" record with combined bibliographic information

<Container>
<DisplayTitle offset="0">Symphonies 5 & 7 [sound recording] : Egmont overture</DisplayTitle>
<Structure label="Symphonies 5 & 7 : Egmont overture / Beethoven">
<Item label="CD 1">
<Div label="Beethoven, Symphony No.5 in C minor, op.67">
<Chunk label="Track 1. I Allegro con brio">
<ContentInterval begin="0" end="432734" mediaRef="IU/MediaObject/4061"/>
</Chunk>
<Chunk label="Track 2. II Andante con moto">
<ContentInterval begin="432734" end="988734" mediaRef="IU/MediaObject/4061"/>
</Chunk>
</Div>
</Item>
</Structure>
</DisplayTitle>
</Container>
Migration process

Mass Store tapes

File Server

MARC/Z39.50

Audio Encoding

Streaming Server

V2 Record Creation

Variations2 Database

Technical metadata

28 & 192 kbps mov

container & media object records

bib & track info

MARC record

wav

28 & 192 kbps mov

bib & track info

MARC record
Initial problems

- Speed of download from tapes in mass store  
  → parallelize the process until CPU is always encoding

- LAME version/streaming problems  
  → needed to find a version of LAME that worked
Migration timeline

Goal: Migrate 10,187 items consisting of 18,899 media files.

- **September 2004**: Initial planning started.
- **Mid-December**: New hardware installed, scripts written; migration begins
- **End of January**: First pass finished; over 1200 items had errors.
- **End of February**: Many problems resolved on second or third pass; 400 remaining problems.
- **March/April**: Spot checking revealed some items appeared to load correctly, but were not playable. We built new validation tools, and identified 200 additional problem items.
- **May 2005**: System goes into production for users
- **End of Summer 2005**: Remaining failures resolved, just in time for fall semester.
Migration progress

Items per Day

- 0
- 50
- 100
- 150
- 200
- 250
- 300

Dates:
- 12/10/2004
- 12/17/2004
- 12/24/2004
- 12/31/2004
- 1/7/2005
- 1/14/2005
- 1/21/2005
- 1/28/2005
- 2/4/2005
- 2/11/2005
- 2/18/2005
- 2/25/2005
- 3/4/2005
- 3/11/2005
- 3/18/2005
- 3/25/2005
- 4/1/2005
- 4/8/2005
- 4/15/2005
Migration: Audio problems

- Files unobtainable from mass store (cache, system down)
- 465 files missing from mass store, but may be available elsewhere (TSM tapes, drawer tapes, redigitize)
- 225 files with unexpected .wav format
- Some poor quality files
Migration: Metadata problems

- Problems that could be fixed in import scripts:
  - Inconsistent values (dates, copyright)
  - Abbreviations
  - Multiple MARC records

- Problems that required manual intervention:
  - 150 Items with offsets beyond end of file (really an audio problem!)
  - 92 Missing MARC records -- some items were never cataloged; some had been deleted
  - 22 Invalid characters in track file or MARC record
  - 4 Items with “dummy” offsets (10 min, 20 min)
  - 2 Items with more than 26 sides
Load testing
New ingest process

- More constrained vocabularies
- Extensive validation
- Item cannot be published until files have completed processing
- Automatic calculation of ending offsets
- Automatic import of MARC record
- Enhanced tools for creating track hierarchy
- “Listen” button allows just-in-time audio validation
Migration Staffing

17 vs. 26 person-months
# of audio players opened daily, beginning of Fall Semester, 2005 (excluding 2 days where logging broke)
User Reaction

- General user response
  - Mostly silence (a good thing!)

- Home use
  - 961 client downloads by unique network IDs as of 11/1/2005 (776 Windows, 185 Mac)

- Good signs of adoption of new tools (e.g. Timeliner, Score Annotation) by students and faculty
Lessons learned

- Start early and have a good plan
- Automate the entire process
- Include validation and allow process to be rerun
- There will be problems with missing or invalid content
- No “one-off” fixes
- Generate good logs and keep them
- Plan for post-migration support
  - There will be cleanup work
Future Plans

- **Variations2**
  - Continue in production
  - Figure out how to fund ongoing support

- **Variations3**
  - Just funded by IMLS National Leadership Grant
  - 3-year project to make a “digital music library in a box” that can be deployed at other institutions
  - Initial partners: Ohio State, New England Conservatory, Bryn Mawr/Haverford/Swarthmore Tri-College Consortium
  - Use own content, but also integrate licensed content from vendors
  - Explore options for metadata creation
For More Information

- [http://variations2.indiana.edu/](http://variations2.indiana.edu/)
- Jon Dunn: [jwd@indiana.edu](mailto:jwd@indiana.edu)
- Ryan Scherle: [rscherle@indiana.edu](mailto:rscherle@indiana.edu)
Disclaimer

This material is based upon work supported by the National Science Foundation under Grant No. 9909068.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.