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





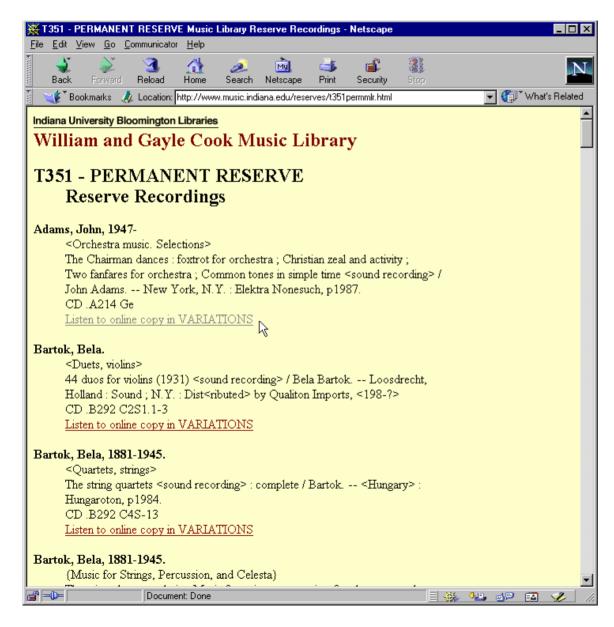
Outline

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

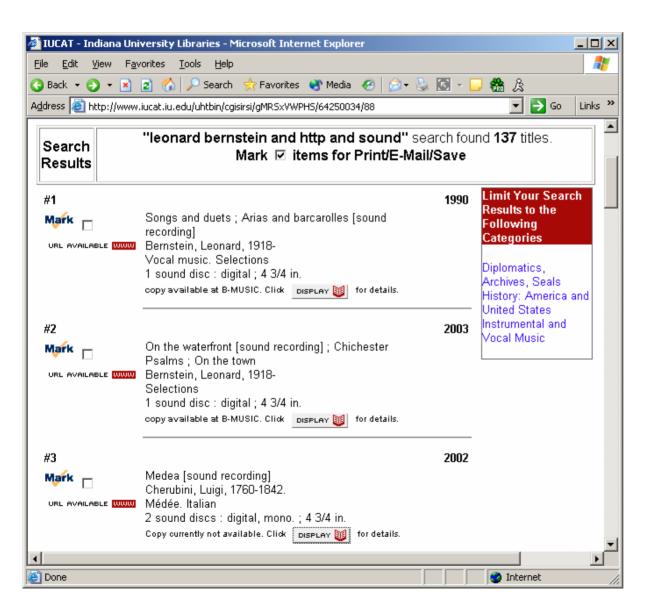


- Digital library of sound recordings and scores in IU Cook Music Library
- Online in April 1996
- 10,000 recordings; 200+ scores
- Access to course reserves, general, and special collections
- Available from workstations in Music Library and selected on-campus locations
- Jointly supported by Digital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

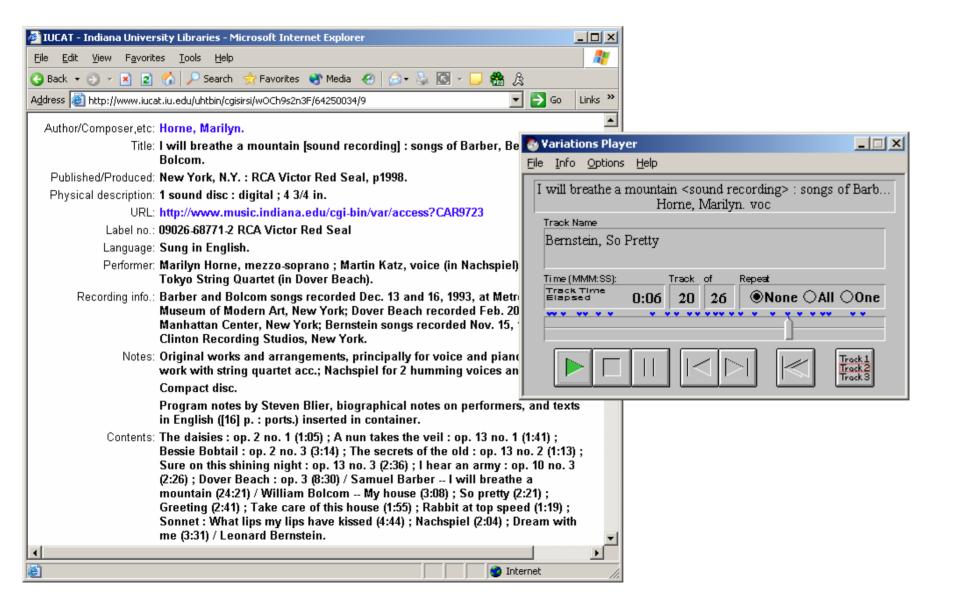




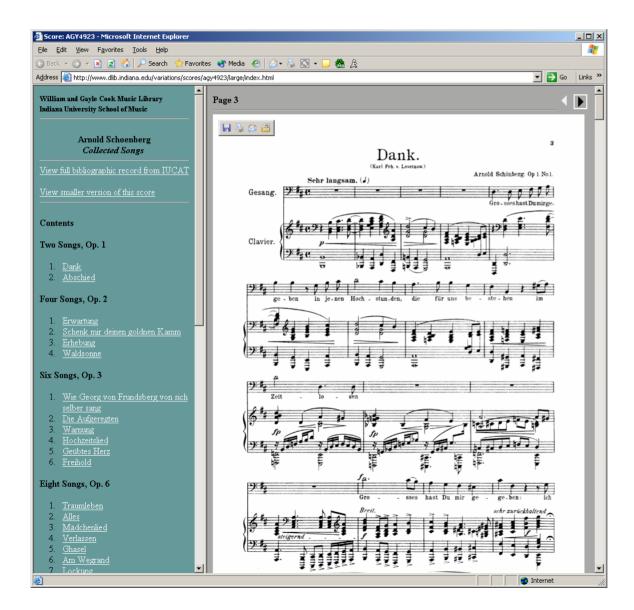














Pre-Variations









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

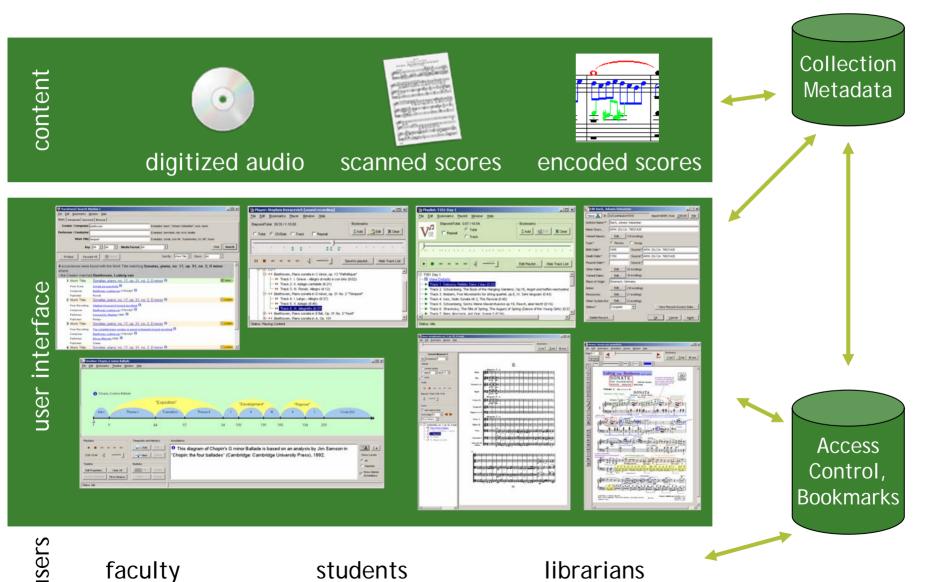


Expanding on Variations

- 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

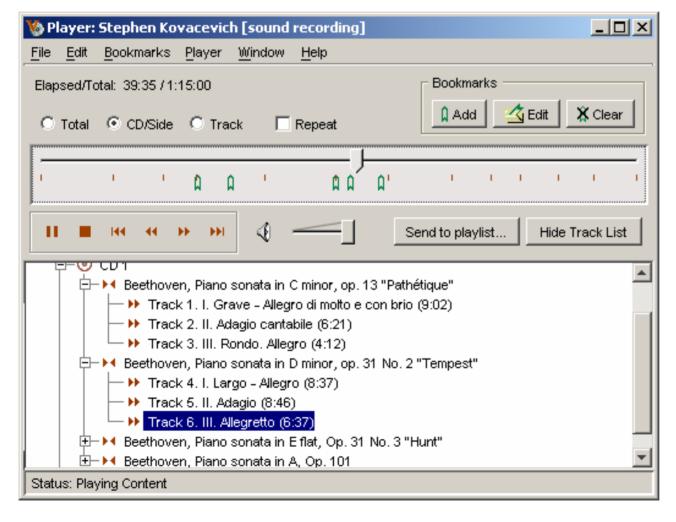


Variations2 Overview





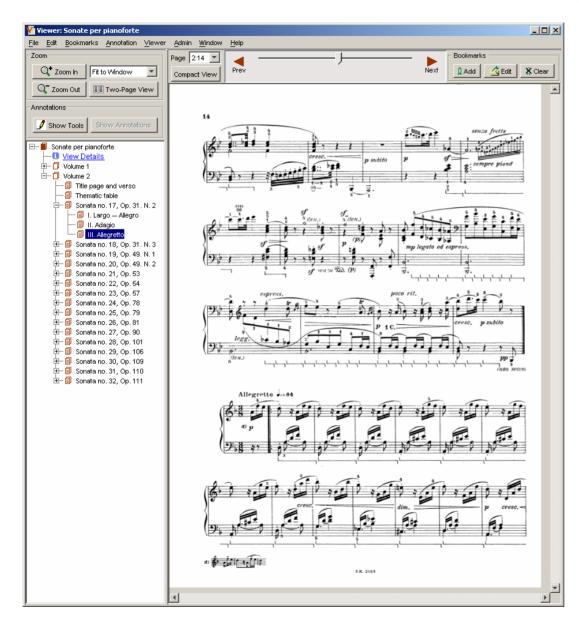
Audio Player



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

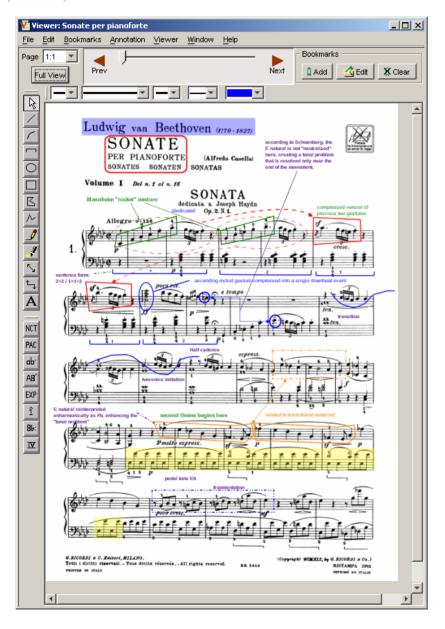


Score Viewer



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



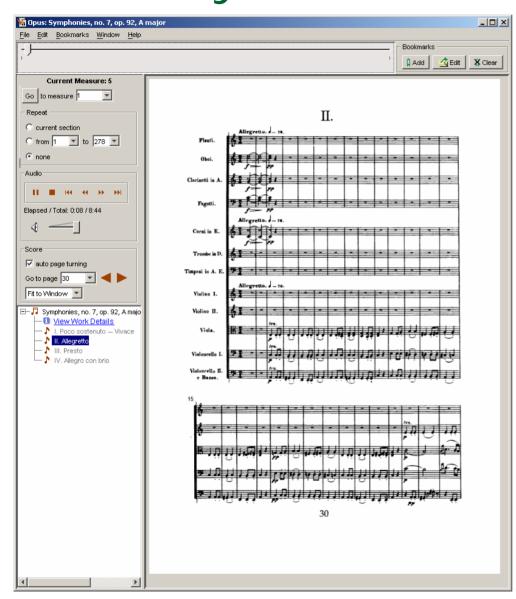


Score Annotation

Users can also annotate scores and save or print the annotations.
Annotations scale with page zooming level.



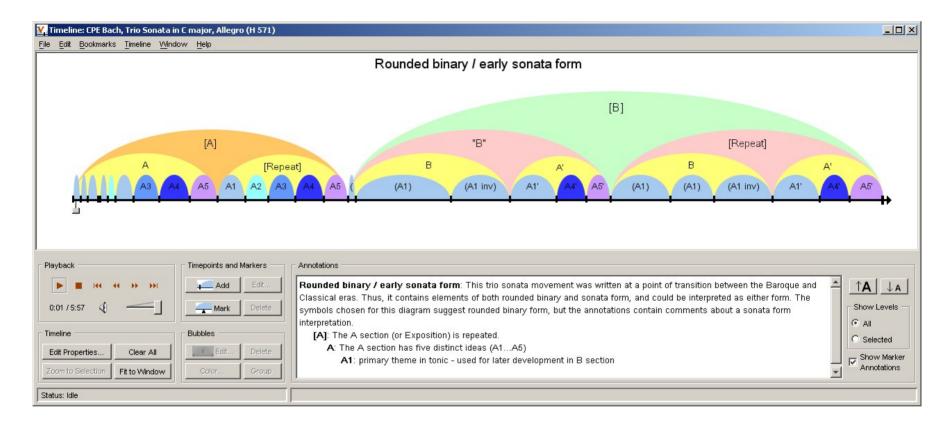
Synchronized Score & Audio



Any recording can be synchronized with any score of the same work for easy navigation and even automatic page turning.



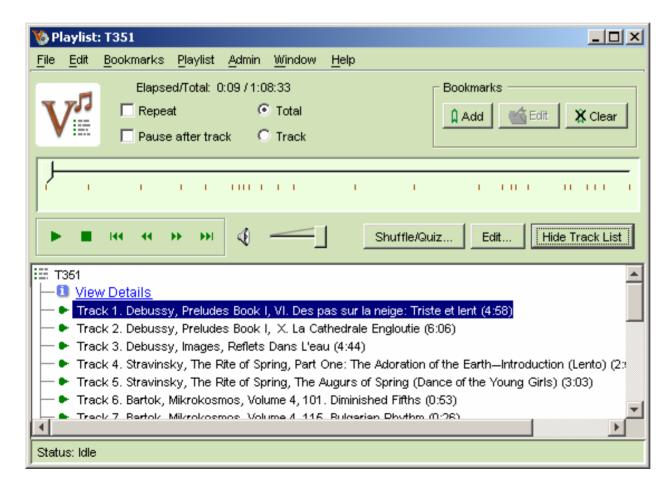
Audio Timeliner



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.



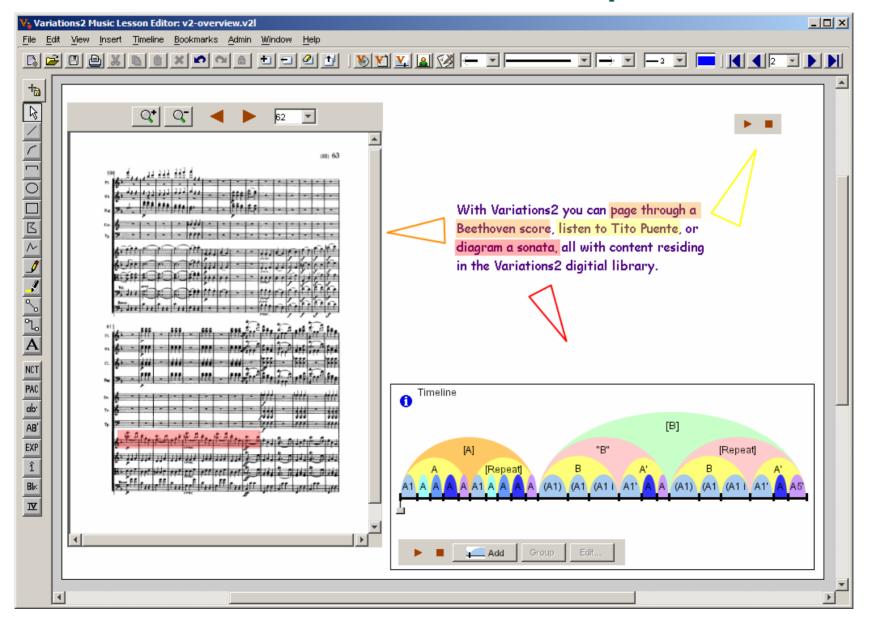
Personalized Playlists



Users can create personal playlists from different albums.

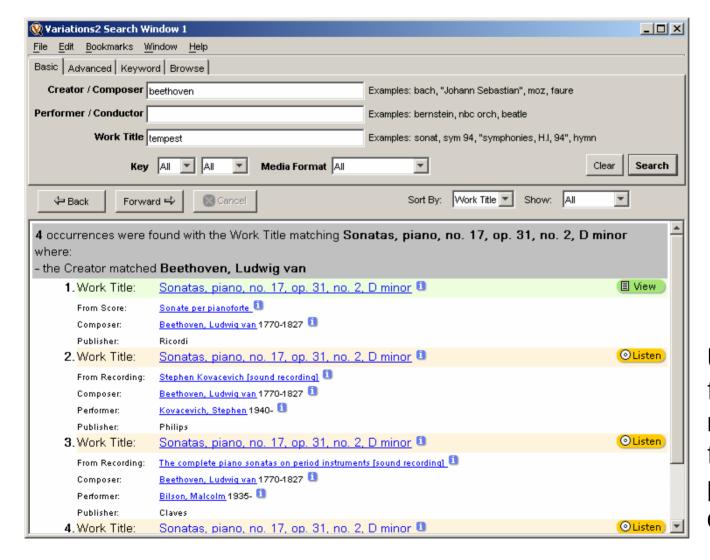


Lesson Editor (experimental)





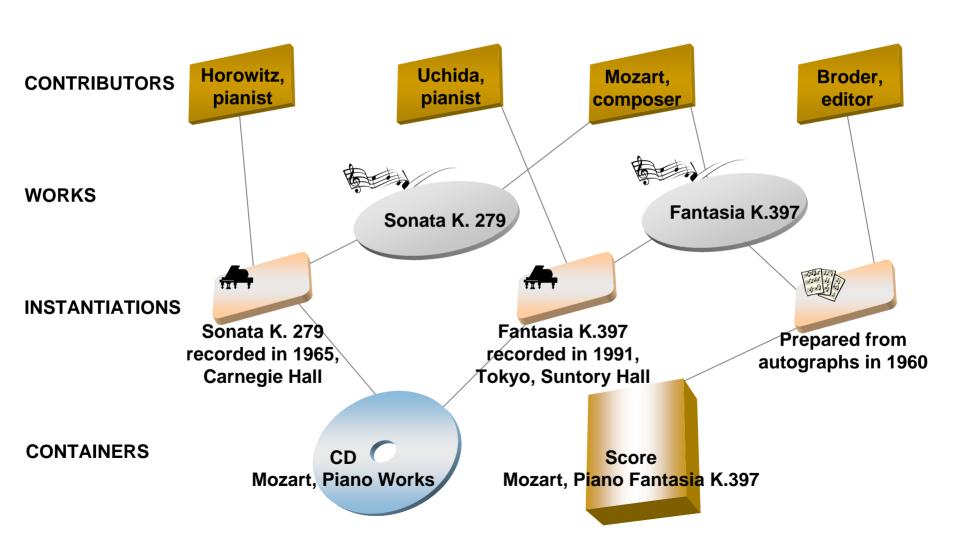
Search Window



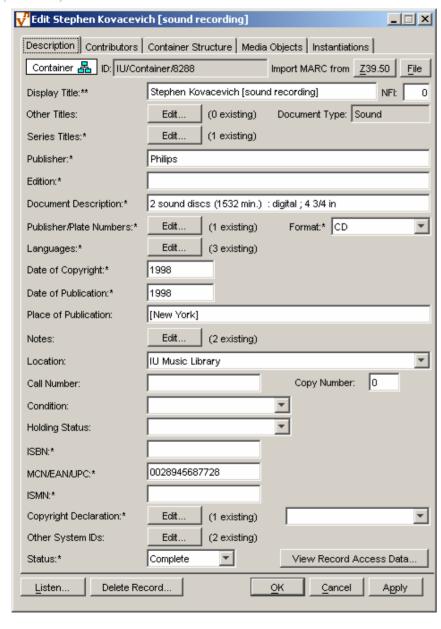
Users can search for music using music-related fields, such as performer, key or media format.



Data Model Example







Cataloging Tools

Catalogers enter metadata for items, contributors, or works. Data can be imported from existing MARC records.



Content and Metadata Comparison

	Variations	Variations2
Master files	CD-quality WAV	
Delivery files	384 Kbps MPEG-1 layer 2 ("MP2")	192 Kbps MP3 28 Kbps AAC
Metadata	MARC + "track file" stored on filesystem	Complex metadata model in DB2



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

- 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



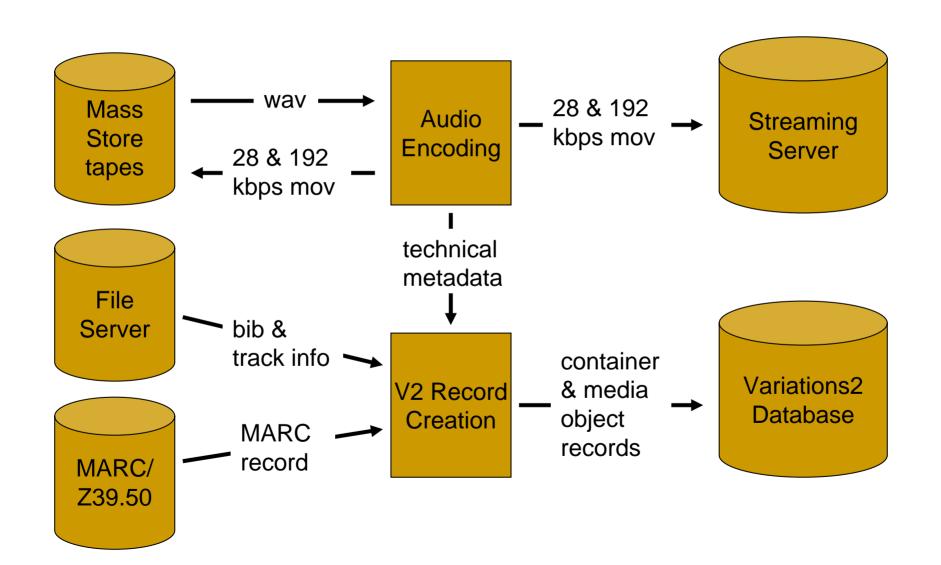
Where we needed to end up

- 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 & amp; 7 [sound recording]: Egmont
   overture</DisplayTitle>
   <Structure label="Symphonies 5 & amp; 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"</pre>
                            mediaRef="IU/MediaObject/4061"/>
         </Chunk>
```



Migration process





Initial problems

- Speed of download from tapes in mass store
 → parallelize the process
 - → parallelize the process until CPU is always encoding
- LAME version/streaming problems
 - → needed to find a version of LAME that worked





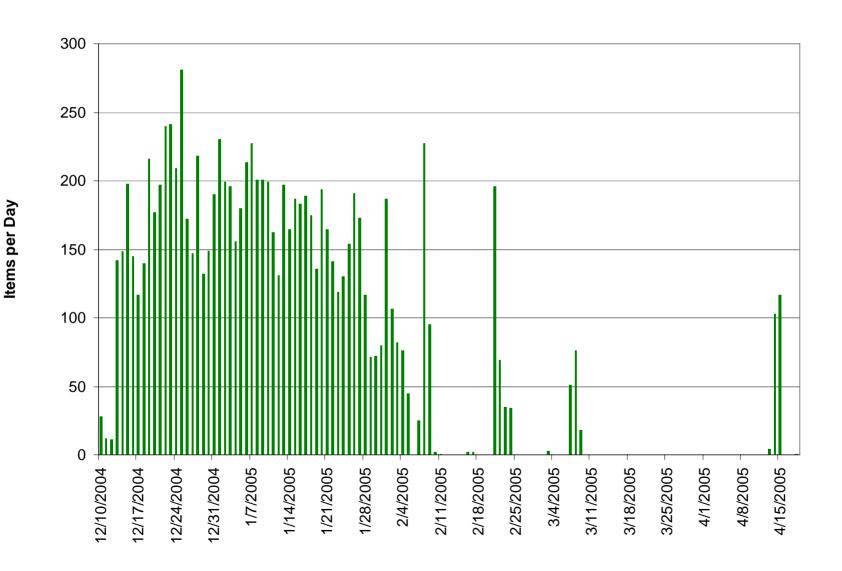
Migration timeline

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





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



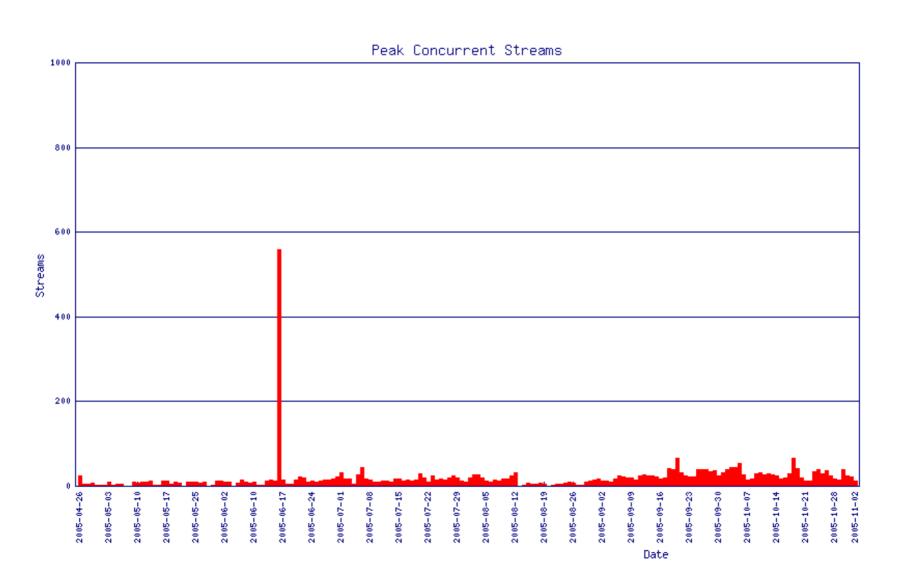


Migration: Metadata problems

- 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



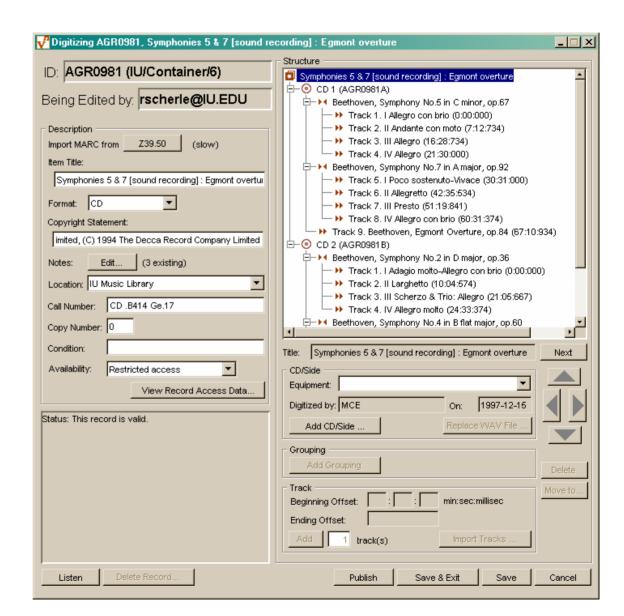
Load testing





New ingest process

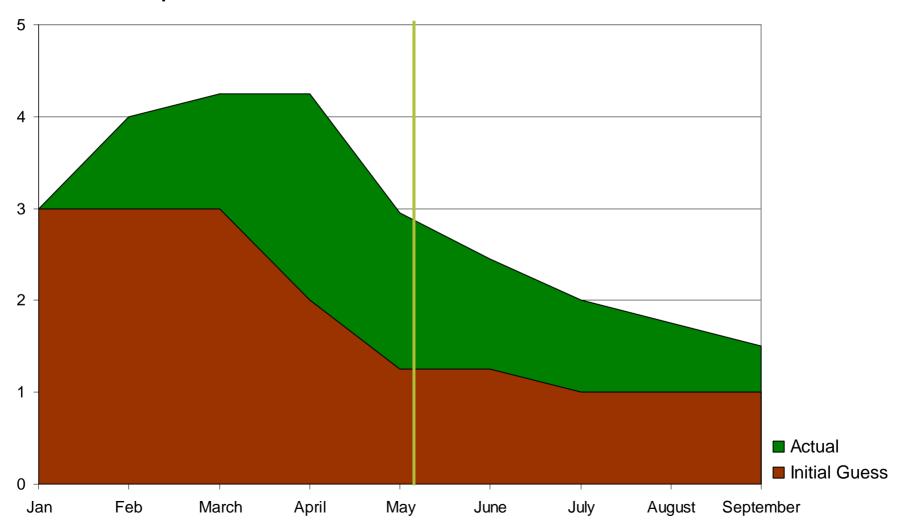
- More constrained vocabularies
- Extensive validation
- Item cannot be published until files have completed processing
- Automatic calculation of ending offsets
- Automatic import of MARC record
- Enhanced tools for creating track hierarchy
- "Listen" button allows just-in-time audio validation





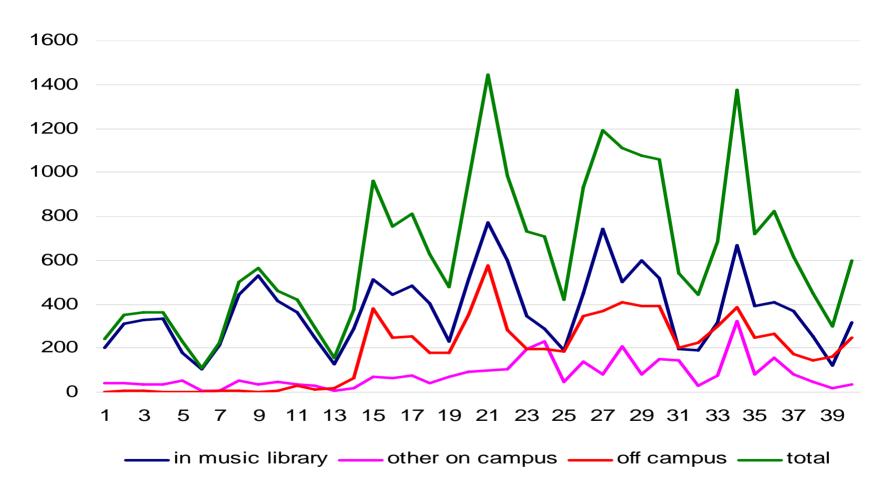
Migration Staffing

17 vs. 26 person-months





Usage



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

- 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/

- Jon Dunn: jwd@indiana.edu
- Ryan Scherle: <u>rscherle@indiana.edu</u>



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.