Peabody Digital Audio Archives Project

ABSTRACT

Goals and major activities

The main goal of the current two-year project is to design and create an effective and economical workflow management system for digitizing analog audio tapes and building a web-accessible digital audio library. Strong emphasis will be placed on using best practice, open standards, and open-source software. As a testbed, the audio tape collection in the Peabody Archives will be used.

Peabody Archives holds about 10,000 hours of tape recordings of concerts and recitals at Peabody over the past forty years. Most of the ensemble recordings are in the public domain. The long-term goal of the Archive is to digitize the entire collection.

We plan to:

- Demonstrate that a very inexpensive server can simultaneously support: audio streaming, search engine and indexing, audio and image digitization, and digital library construction.
- Develop and design economical workflow management for audio tape ingestion using appropriate procedures required to digitize aging analog audio tapes
- Digitize about one-third of the collection to test the framework and its scalability
- Clarify the copyright issues of the Archive
- Collect user feedback and assess usability
- Evaluate the feasibility of using an inexpensive server as foundations for search and distribution of terabyte-scale audio data
- Develop audio segmentation software to automatically separate pieces and movements within a piece
- Use an open-source software Greenstone to create a flexible digital audio library
- Enhancements of Greenstone:
  - Evaluate existing audio metadata formats (e.g. MPEG 7/21, MusicBrainz) and extend Greenstone if applicable
  - Evaluate emerging web-services such as SOAP, UDDI, WSDL for digital libraries and extend Greenstone if applicable
  - Evaluate other digital library model such as FEDORA and modify Greenstone if applicable

Anticipated results

- Suite of open software for establishing digital audio libraries
- Tested prototype of workflow management system for digitization of analog tapes and building of digital audio library
- Establishment of a large audio collection of Classical music, including access to rarely recorded solo repertoire for music students and teachers
- Historical studies of performance practice and repertoire of a Conservatory
- Important database for scientific studies in music information retrieval, expressive performance, tempo tracking, timbre recognition, performance style analysis, etc.
- Offer economical analog tape digitization service for similar projects
National Impact

This project addresses the national impact by developing and demonstrating an efficient and effective workflow management system for building a digital audio repository from an analog tape collection. Both the digitization workflow system and the resulting repository should serve as a model for other libraries with similar holdings to build more accessible audio archives.

With recent decrease in hardware costs, availability of high-speed networks, and new compression technologies, access to large amount of audio and music, in particular, is becoming a reality. Yet the knowledge of how to build and design a collection of music in digital format is still in its infancy. Also, as the analog tape preservation copies in various institutions begin to deteriorate there will be an increasing demand for both preservation and access to these collections.

The access to school concert archives will have tremendous educational value for music students as it allows them to listen to other students and for many instruments, commercial recordings for solo repertoire is either very limited or non-existent.

The audio segmentation software (see Design) should also be useful for other types of recordings where a number of different pieces is contained on one tape, e.g. concerts of any type of music and radio shows.

With the practical knowledge gained and acquisition of equipment, Peabody should be able to offer high-quality and economical digitization service for libraries, music schools, and other organizations with analog tape collection.

This project addresses four of the IMLS priorities for 2002 Research and Demonstration projects:

- By providing Internet access to recordings of Classical music, this project will “help libraries take a leadership role in the education of lifelong learners in the 21st century.”
- By using hardware that became recently affordable, using open-source software, and novel approach of tape digitization process using pipelining and parallel processing, this project will “enhance library services through the effective and efficient use of new and appropriate technologies.”
- By developing web-based search engine and high-quality compressed audio file distribution system, this project will “enhance the ability of library users to make more effective use of information resources.”
- By conducting usability studies of the resulting digital audio library collection, this project will “address how people find information, how they would prefer to find information, and how they benefit from information.”

The project also addresses issues cited in the Recommendation 1 of the President’s Information Technology Advisory Committee’s Panel on Digital Libraries in Digital Libraries: Universal Access to Human Knowledge (February 2001): “Support expanded digital library research in metadata and metadata use, scalability, interoperability, archival storages and preservation, intellectual property rights, privacy, and security, and human use”
Adaptability

- Digitization workflow management system will be applicable to other analog tape digitization projects and can be adapted to vinyl disc digitization projects.
- The process of building digital library repository will be applicable to other audio collection including digital audio (e.g. CD and DAT) collections.
- Separation software should be adaptable to other types of concerts (rock, jazz, folk) or any other single audio file containing multiple works, such as radio shows.
- Metadata extensions will be applicable to a wide variety of audiovisual archives.
- Web-services extension will be adaptable to digital library collection in general.
- All software developed will be available as open-source software for ease of adaptability.
- Web documentation designed for so that the workflow management and digitization process can be easily replicated.
- The productivity and problems encountered will be documented in detail as well as offering advice and tips based on our experience.

Design

The search engine and basic building of the collection will be based on Greenstone, an open-source software. One of the goals of the project is to extend the software according to the guidelines and best practice that are established in the audio-visual digital library community. Search engine will have free text search and Boolean searches on metadata including: title, movement title, instruments, performers, composer, composition date, and concert location and date.

The overall design of the project can be divided into two phases: preparation phase and production and maintenance phase. Phase I will involve preparation of the digitization process and creating the infrastructure for the building of the digital collection. Phase II will involve the digitization of the analog tapes and concert programs and building of the digital collection.

Phase I:
Several utilities will be developed, in order to reduce the cost of audio digitization, in most cases, to reduce human intervention, which is always expensive. Also, whenever possible, existing open-source software will be searched and used as the basis for all of the software development.

Production software
Two principle ways to reduce the production time, thus cost, are parallel processing and pipelining. Both of these methods are incorporate in our production system.
By using an eight-channel analog-to-digital converter, four different stereo tapes can be digitized at the same time. Refer to Figure 1 for the following discussion. Furthermore, since each tape is relatively long (typically thirty to forty minutes), one person should be able to operate two machines and scan the corresponding concert program notes (typically one or two pages). Therefore the following process will be repeated in the production cycle:
1. Mount tape on the first tape machine and start digitizing
2. Scan the concert program notes
3. Mount another tape on the second tape machine and start digitizing
4. Scan the concert program notes

Obviously, the tapes must be unmounted after the initial cycle. Also note that each tape must be rewound (that is how the tapes are stored) before it can be played back for digitization. The fifth tape machine in Figure 1 is used for making new analog preservation copy for certain selected tapes. Although this project is not aimed at creating preservation copies, it was decided that, for some of the older and valuable tapes, preservation copies should be made while digitizing. Since this will be done at the same time as the digitization, the cost for this process is minimized except for the cost of the tapes.

In order for the production system described above to work, specialized software must be developed that will control the recording on the server, remotely from the operators’ workstations. This software will provide a user-friendly graphic user interface on the workstation which can be used to start and stop the recording process and also to enter minimal text information, i.e. tape identifier. The control signal and text data will be sent over the Ethernet connection between the workstations and the server using TCP/IP protocol. This program will also alert the operator if an error, such as clipping, occurred during the digitization or that overall recording level was too low so that the appropriate measures can be taken to ensure high quality of the results. The workstation will also run the scanning software to control the scanning process and the resulting image will be sent over the Ethernet to the server to be OCR’d.

**OCR software and post-processing**

It is anticipated that a standard commercial OCR software will be able to process the concert program notes using specialized dictionary containing composers’ names and names of movements (e.g. adagio, allegro, largo). A post-processing utility program will be written to recognize name of composers, titles
(e.g. sonata, minuet, symphony), movements, which will reduce the metadata creation time. Also automatic name authority control will be implemented to recognize variant spellings [Tim: help me out here.]

**Audio segmentation software**

Because each tape will contain several different pieces, automatically separating the pieces will eliminate the time-consuming process of manual separation. The strategy that will be used for this task is to train the computer to recognize applause which normally separate each piece in a concert. A more sophisticated tactic must be deployed to separate pieces in a song recital where several pieces are often sung between applause. The solution to this situation can be aided by formatting of the concert program where larger spaces are inserted between pieces where the singer expect the audience to applause along with the technique used to separate movements within a piece. Recognizing the separation between movements is anticipated to be a much more difficult task. Some features that can be used for separation include long silences, sudden increase in coughing in audience, and tuning of instruments. Although a long pause within a movement and *attacca* (where two consecutive movements are performed without a break. When the software is unable to segment movements that matches the movements specified in the concert program, an error message will be generated to indicate that manual intervention is required.

A software with a graphic user interface will be developed, based on existing open-source software, which facilitates the manual separation of movements. This software can also be used to check the correctness of the automatically segmented audio files.

**Metadata extension and conversion software**

Extending or provide conversion to other metadata formats such as MPEG 7/21, MusicBrainz, MARC, and Z39.50.

**Metadata distribution and remote access software**

Appropriate implementations of SOAP (Simple Object Access Protocol), UDDI (Universal Description, Discovery, and Integration), XSLT/XQuery, WSDL (Web Services Description Language), etc. to extend Greenstone/CORBA. Register with OAI. Uploading to OCLC, RLIN?

**Network access simulation and monitoring software**

In order to estimate the potential and maximum capacity of the server and the network, network access simulation and monitoring software will be developed to simulate multiple search requests and audio streaming requests. This will be important in providing quantitative measures for the evaluation of the system.

**Phase II:**

This phase will involve the actual digitization of audio tapes, scanning of the concert programs, OCR, and ingestion of both audio data and scanned images and metadata creation. It will also include maintenance of the software and quality control of the digital collection. Furthermore, approximately 500 hours of older and important tapes selected by the Archivist will be transferred to analog tapes at 15ips as a preservation copies.
Two sets of digital tape backups of the preservation copies will be made and stored in fire-proof boxes. One box will be stored offsite. High-quality MP3 audio files along with 30-second incipit will be placed online.

Scanned two copies of concert programs (TIFF) will be printed on 600 dpi printer on high-quality acid-free paper and bound. One copy will be placed in the library stack for general use and the other will be placed in the Archives as a preservation copy. Compressed version (JPEG or DjVu) will be placed online.

**Usability tests**
The usability assessment of the Web-based search engine will begin with a heuristic evaluation of a prototype of the search interface. This evaluation will include a review for compliance with accessibility guidelines. When the tapes have been digitized, test users will participate in a scenario-based, think-aloud protocol. The results of this test will inform refinements to the search interface. A second round of scenario-based tests will complete the usability assessment. Test participants will represent target user groups, including music students and teachers.

**Management Plan**
Ichiro Fujinaga, as Principal Investigator, will be responsible for the overall management of the project. He will closely work with and supervise the programmer. In the second year, he will be responsible for debugging, maintenance, and improvement of the various software and system administration.

Bill Racine, as the Production Manager, will manage the digitization involving Recording Arts students. He will maintain the production schedule and collect statistics of productivity that will be used for evaluation of the process. He will also be responsible for selecting tapes to be baked ad for quality control of digitized audio and that of analog preservation copies.

Elizabeth Schaaf, as the Archivist, will select the tapes to be digitize and select tapes for which new preservation copies are to be made. She will coordinate the transfer of the tapes from the Archive to the Production rooms to ensure security and safety. She will be responsible for monitoring the quality of the metadata, and the long-term maintenance of the web site.

Tim Dilauro, as the Consultant, will be consulted by PI on the issues of workflow management, collection building, metadata, and web-services. He will also work with the programmer.

Teal Anderson, as the Usability Specialist, will perform a heuristic analysis of the web-site design and the search engine interface. She will also conduct two usability tests of the collection.
Tasks and dates:

10/2002  Install Greenstone; Purchase and install A/D converter, server and tape backup system; Setup OS and implement backup procedure
11/2002  Develop multi-track recording and playback system; Develop TCP/IP-based remote control of the recording system.
12/2002  Batch audio format conversion software
01/2003  begin scanning and digital transfer of CD-data to create test data
02-03/2003 OCR test; GAMERA, automatic authority name control; rights management utility
04-05/2003 Develop of audio file segmentation software; Start digitization
06/2003  Metadata extensions and conversions
07/2003  Network monitoring software; Network access simulation software
08/2003  Go online, Web services extensions improve OCR system;
09/2003  Debugging, final reports on software, First usability test
10/2003-08/2004: Digitization, quality control, software debugging and improvements
08/2004  Second usability test
09/2004  Final assessments and reports

Here is the summary of major tasks assigned to the programmer and the students:

Programmer

- Setup and maintain a server and PCs
- Setup tape backup system
- Setup project web site including questionnaire for users
- Install soundcard and A/D converter
- Install and learn to use and program Greenstone
- Management of digitally-born audio
- Software development and documentation
  - Implement OCR for metadata creation
  - Batch conversion of audio formats (MP3)
  - Automatic authority name control
  - Automatic rights management utility
  - Access simulation and network monitoring
  - Metadata extensions and conversion
  - Web service extensions
  - Audio file segmentation

Students

- Prepare tapes for digitization (cleaning, baking)
- Digitize tapes
- Make preservation analog copy of selected tapes
- Digitize associate documents and initiate OCR
- Enter minimal identifiers for audio and text digitized
- Maintain tape machines
Budget

One of the major goals of this project is to reduce the costs of digitization process and digital library construction.

The major equipment purchases include a server with over 400GB of disk space ($2500), 200GB/tape digital tape backup system ($3500) and a 8-channel A/D converter ($6000). Of the five Sony 5002 tape decks, two are in working order but the other three need to be refurbished. Other hardware will be provided by Peabody. The digital tapes will be used as temporary preservation copies and also for daily backup of the server. Analog tapes will be used for making new analog preservation copies of older and important tapes.

The students in Recording Arts program will be effectively used to perform the time-consuming task of digitization, totaling 3000 person-hours to process over 3500 hours of music.

An experienced programmer will be hired for the first year of the project to further reduce the cost of digitization by developing software that will automate the ingestion process.

Since the PI is on a 9-month contract, he will be compensated for 1/9th of his salary so he can work on the project over the summer months. Two experts from the Digital Knowledge Center of the Johns Hopkins’ Eisenhower Library will be hired. Tim Dilauro, a digital library expert will contribute 5% of his time as a consultant and Teal Anderson, a usability specialist, will contribute 2% of her time conducting usability tests. The archivist Elizabeth Schaaf and the production manager Bill Racine will contribute 15% and 12.5% of their time, respectively.

Contributions

Contributions from Peabody include 20% of the faculty member during the academic year, 15% of the archivist and 12.5% of the digitization manager. Peabody will consign five Sony professional tape machines, originally purchased at $22,000 each, a noise reduction unit and a multi-channel sound card. Tape testing equipment and tools for aligning tape heads and general maintenance of tape machines, audio playback systems for quality control will also be made available. Peabody will supply 100 highest-quality analog audio tapes for the preservation copies. In addition to the equipment listed in the budget, two other high-end PC workstations will be made available for software development and two flat-bed scanners for document digitization. Although not in the budget, there will be other institutional support from our existing infrastructure of hardware, software, and network access and support.

Personnel

Ichiro Fujinaga

Has considerable experience in a variety of areas suitable for this project. He has 28 years of system administration experience and is an accomplished programmer. He has designed and built the world’s largest Handel discography database now available at McGill University (URL). In the fall of 2001, he has recently given a seminar on Music Data Mining surveying music search engines, metadata formats ad web services protocols. For the last two years, he has been an co-Principal Investigator for NSF DLI2 funded Levy Project, which involves design of workflow management system for digitizing sheet music and developing a flexible optical document recognition software called, Gamera, which is based on PI’s
optical music recognition program. He has also been developing a system for recognizing orchestral instrument recognition system, the experience from which will be valuable for developing the audio segmentation software

Elizabeth Schaaf
Experience archivist
Long-term archivist of Peabody Audio Archive

William Racine
Experienced recording engineer
Familiarity with the Peabody Audio Archive

Timothy DiLauro
Expert in workflow management, digital library design, metadata, web-services

Teal Anderson
Usability expert

Project Evaluation
Quantitative assessment of the output and outcome-based assessment will be used to evaluate the effectiveness and usefulness of the digitization process and the digitized collection.

- Two usability testing involving music students and teachers will be conducted
- Web access statistics both locally and over the Internet will be monitored, tallied, and reviewed:
  - Analyze visits to different pages within the site, such as search engine page, FAQ page, concert program images, etc.
  - Average length of streamed audio
  - Metadata search statistics
  - Remote access request for web services.
- As part of long-term outcome-based assessment, automatic log analysis will be maintained and displayed on the project web site.
- Quality of audio, both the conservation copies and the compressed format will be assessed.
- The project web site will include a questionnaire asking users to comment on the search engine, general site setup, audio quality, etc.
- Production output:
  - numbers of concerts (hours) digitized
  - numbers of hours used for audio digitization
  - number of concerts (hours) processed for preservation replacement copies
  - number of hours used for the preservation process
• number of pages of concert programs digitized and OCR’d.
• number of hours used for digitizing the concert programs
• number of hours used to correct the OCR results

• Digital library statistics:
  • number of pieces in the metadata
  • number of pieces (movements) and made fully available in the public domain
  • number of pieces made available as 30 second incipit
  • total number of hours of streamed audio

**Dissemination**

Dissemination will begin in the earliest stages of the project and the results will be made available worldwide.

Informative project Web site with links from both the Peabody Library page and Johns Hopkins Library page. This site will include the description of the project, technical reports, tips, contact email address, search engines, bibliography, and links to other sites. All of metadata ingested in this project will be freely available on the Internet along with audio files that are in the public domain, and thirty-second incipit of every movement in the collection. Software will be available through sourceforge.net, where there will be project web-page, mailing list, the source code and the CVS. Information about the project will be disseminated through various mailing lists (e.g. MIR, Greenstone).

Technical reports to be generated and made available on the project Web site include:
• Software development
• Workflow management system (data workflow, staffing, metadata creation, and quality control procedures)
• Production report, including cost estimate
• Usability test report
• Digitization advice
• Metadata for music files
• Web services for digital libraries

These reports will be combined to be presented at conferences (e.g., JCDL, IMSL, ISMIR, ICMC, MLA) and published in conference proceedings and articles will be published in leading professional and technical journals (e.g. D-Lib, JASA)

**Sustainability**

Peabody has been committed and will continue to maintain and support Peabody audio archives. Substantial portion of archivist on-going task is to maintain the Peabody audio archives. With the implementation of ingesting current and future Peabody concerts the collection will grow and kept up-to-date. We will aggressively seek funding for digitization of the rest of the collection using the workflow management system and software developed with this project. The entire collection will be
duplicated on suitable media for long-term preservation when the appropriate technology emerges. The web site and software developed during this project will be made available indefinitely.

The usefulness of this collection for various applications including use of this collection as testbed for research will also promote long-term availability of this collection. The PI is interested in integrating musical scores to audio so that the users can listen and look at the scores at the same time.

DKC is also interested in sustaining and maintaining the collection and future research. Music information retrieval community is desperately looking for a test bed for audio database.

Finally, one of the major goals of this project is the use of the low-cost hardware. This facilitates the long-term sustainability, not only for this project but other similar endeavors, because they will be inexpensive to maintain.