



Learning on the Fly: Large-Scale Audio Digitization of the Texas Senate Tapes

Best Practices Exchange

November 14, 2013



Overview



Texas Senate Tapes

- 26,000 cassette tapes, most 45 minutes per side
- Recordings of Senate committee hearings and floor hearings between 1972 and 2006
- High reference use and duplication requests
- Often the only record available
- Few transcripts exist, especially for the early years
- Approximately 260 known damaged tapes



Preservation Issues



Repeated use and improper playback resulted in damage like this to many of the tapes.

Other factors:
Age
Deterioration
Quality



Previous Digitization Efforts

- Audio digitization workflow was in place from 2003 – 2006
 - earliest tapes digitized for patron access, not preservation
 - utilized a Sony TC-KE400S stereo tape deck, an unidentified sound card in a TSLAC computer, Gold Wave software, and storage on TSLAC servers
 - established metadata collection procedures
 - MP3s and a simplified finding aid for public access



LSTA Funding

- Late 2012 – large amount of Library Systems Technology Act grant funds became available
- Must be used in accordance with the statewide plan
- State plan included digitization of archival materials for public use (Goal 1, item 3)
- All funds to be expended by August 31, 2013



Project Goals

- High quality preservation copies
- Manageable digital copies for research use



Project Considerations

- Accelerated timeline
- Previous costs analysis provided foundation
- Procurement rules and requirements
 - Council on Competitive Government
 - Department of Information Resources
 - Commission approval for contracts over \$100K
 - Bid requirements
- Storage



RFP Development

- Staff gathered additional information on digitization specifications
- RFP review team assembled, with agency staff from the Archives, Purchasing and IRT
- Identified all known requirements and deadlines (internal and external)
- Request for Proposals draft submitted to Purchasing for review on January 22
- Review criteria established



RFP Scope

- 1 digital preservation file in Broadcast Wave format per cassette tape side, which will result in 2 files per cassette tape; and
- 1 digital access file in MP3 format per cassette tape side, which will also result in 2 files per cassette tape
- Standard XML preservation metadata for each the digital files created
- Structural, descriptive, administrative, and technical XML metadata for the digital files.



RFP Award

- Proposals received from four vendors
- George Blood Audio and Video selected
- Contract details clarified included:
 - criteria for the packing and shipping of the tape
 - digitization specifications and final product
 - hardware used
 - timeline for review and corrections



RFP Award

- Contract for Services #306-13-8343
- Total project cost: \$289,676.80
- First outbound shipment on or before April 24
- Second shipment on or before June 3



Project Management



Packing and Shipping

- Tapes sent in two shipments – late April and early June
- The vendor sent two staff members to pack the tapes
- Tapes shipped via a fine art shipping company in a refrigerated truck to the vendor in Philadelphia
- Each shipment was insured



Inner box with close fit for tapes





Inner box stacks neatly in outer box





Quality Control



- Vendor's staff randomly listened to 3% of the tapes during their validation process
- TSLAC's IRT staff ran the files through a validation process, comparing metadata sent by the vendor to our internal descriptive metadata



TSLAC's IRT Validation Process

- Mount batch of files received onto a Linux server where a script can be run to:
 - pull required information out of the media files
 - get a list of warnings and errors about the formatting of the mp3, wav, and xml files
 - verify existence of required files for each cassette side
 - check to make sure vendor is providing specified file structure and access required.



- Make backups of all vendor files
- Run a script that compares a md5 hash of a given backup file with the md5 file submitted by the vendor. This lets us know if the files are identical or if some corruption has occurred in the backup process.



- Once all files and correction have been validated and backups made, run another script that does an inventory check that compares a list of file names received, to our in-house inventory of cassette names that adds a spreadsheet entry about any missing files.



- TSLAC validation process produces a spreadsheet with the following data:
 - File name
 - Size of file (bytes)
 - Playtime of the file
 - Bit rate
 - Sample rate
 - Error messages



Error Messages

- File should probably be padded to nearest WORD boundary, but it is not (expecting 417103870 bytes of data, only found 417103869 therefore short by 1 byte)
- Chunk (data) size at offset 40 is zero. Aborting RIFF parsing
- Unknown data before synch (ID3v2 header ends at 7640, then 128 bytes garbage, synch detected at 7768)



TSLAC Quality Control

- Once validated, two Archives staff members listen to 10% of the tapes, checking for clarity, blank spaces, or other issues.
- A list of problems found is compiled and sent to the vendor for correction.



- Problems found through listening to the digital copies were few, such as:
 - plays at very low volume for last half, almost unlistenable
 - audio periodically cuts in and out
 - patch of no content (audio hiss/hum) in middle
 - tape speed too fast at start (0:00 to 4:30)
 - severe audio buzz throughout



- Almost all of the problems were due to the condition of the original tape and could not be fixed
- Vendor listened to all tapes with reported problems and was able to improve the quality on a handful
- Slides from vendor show some solutions



Tape is now playable after hand-winding





Vendor example of
“whatever it takes”:

by shifting the shell
slightly during
playback, tapes with
lubricant failure can
play because they
touch fewer guides.



Storage and Hardware



- Digital files received
 - MP-3 files for research use
 - Broadcast wave files as preservation master
 - XML file
- Copy of the MP-3 files will be stored on networked server and made available for research use through TSLAC website.
- Master MP-3 and broadcast wave files stored on four 12 TB external hard drives with RAID 5 configuration.



Backup Copies

- External hard drives are used to store two sets of backup copies.
 - Backup 1 – files on the RAID drives will be mirrored onto four identical RAID external drives for the first backup copy.
 - Backup 2 – files from the RAID drives will be backed up on external 4-terabyte drives, using 2 drives per RAID drive.



Metadata



- Descriptive metadata compiled from the tapes and their cases by Archives staff prior to shipment
 - File number
 - Date
 - Legislative session
 - Committee name
 - Legislative bill numbers
 - Existence of a transcript
 - Subject headings (infrequent)
 - Notes



Vendor Metadata

- The vendor sends technical metadata about each shipment
- The technical data will be maintained for documentation but not added to TSLAC's public web interface
- The web interface will contain descriptive metadata in a searchable application.



Technical Metadata

- file name
- original file format (audiocassette)
- audio bit rate
- audio sampling rate
- sample rate stock length
- stock manufacturer
- cassette base material
- sound field (mono/stereo)
- engineer file check list
- number of audio channels
- speed
- capture device (numerous fields about device used)
- host computer
- operating system (OS), etc.



Output for Research Use

- MP-3 files will be made available for streaming or downloading online
- Broadcast wave files will not be online, but copies will be available upon request.



Lessons Learned



- More than three months planning time is best to adequately plan a large digitization project
- Clearly specify your computing requirements (ex. PC vs. Mac)
- Contract with the vendor from the beginning to provide both a master and use set of files



- Build in time to work within your IT infrastructure and state purchasing and IT requirements.
- Determine prior to the start of the project how the materials will be stored and accessed
- Secure your funding for storage prior to the start of the project
- Prepare for the unexpected



Contact Information

Jelain Chubb

State Archivist and Director,

Archives and Information Services Division

jchubb@tsl.state.tx.us

512-463-5467

Laura Saegert

Assistant Director for Archives

lsaegert@tsl.state.tx.us

512-436-5500