“Accessing the Texas Digital Archive”
Long-term Digital Preservation and the Texas Digital Archive

E-Records Forum
Nov. 4, 2016
Digital Preservation in 30 minutes or less (probably more) . . .
WHAT IS A RECORD?

Texas Code 441.108(11) defines a “state record” as:

- “means any written, photographic, machine-readable, or other recorded information created or received by or on behalf of a state agency or an elected state official that documents activities in the conduct of state business or use of public resources. The term includes any recorded information created or received by a Texas government official in the conduct of official business, including officials from periods in which Texas was a province, colony, republic, or state.”
Electronic Records

• Sec. 441.189. ELECTRONIC STATE RECORDS.
• (a) Any state record may be created or stored electronically . . .
Characteristics of an Electronic Record

• Four essential characteristics:
  – **Authenticity**-A record must be what it purports to be.
  – **Reliability**-A record must be a full and accurate representation of the transactions, activities, or facts to which it attests.
  – **Integrity**-A record must be complete and unaltered.
  – **Usability**-A record must be able to be located, retrieved, presented, and interpreted.

• Digital preservation is the challenge of maintaining all of these characteristics over time
Authenticity = TRUST
Reliability = completeness
Integrity = Protecting the bits

Example of Bit Rot
Usability
It’s all about the metadata!
(Sung to the tune of Meghan Trainor’s – “All About the Bass”)
Authentic Electronic Records

• It’s all about the Metadata: documents the activities of creation and use
  – Administrative metadata:
    • Access information
    • Audit trails
    • Retention Schedules
  – Technical metadata:
    • File format information
    • Checksums
    • System information/requirements
  – Descriptive Metadata
    • Bibliographic data (Who, what, when, where, why)
    • Indexes/Finding aids
• Unlike some paper records, digital records do not survive without constant attention

• Biggest Challenge facing electronic records - **CHANGE**

**Preservation Challenges**

- **Technical obsolescence**
  - Hardware, software, media

- **Media Deterioration**
  - Magnetic, Optical,

- **Technological Dependency**
  - Hardware, software, etc.
Preservation Challenge

- **Short-term (0-5 years)**
  - Highly active – on-line storage
  - “Normal” management issues
  - Possible application version upgrade

- **Mid-term (5-10 years)**
  - Less active – Near-line/Off-line storage
  - Multiple version control issues
  - System upgrades
  - Possible hardware/software migration

- **Long-term (10+ years)**
  - Least active – Off-line storage
  - Migration/conversion likely

Eventually all electronic records must migrate
When Records Are at Risk

• Software is obsolete or getting to end of life
• Dependency on specific hardware or system
• Non-standard/specialty formats used
• Systems being replaced
• Records no longer in use
• Records are on removable media
Get control over your records!

• Starts with Records Management
  – Records Retention Schedules
    • Identify what records you have and what purpose they serve
    • Establish appropriate retention periods to manage volume
      – Dispose of records at the right time
      – Protect records that need to be retained
  – Recordkeeping systems
    • Organize and Categorize Records
    • Centralized control
    • Improved Access and storage

– Policies and Procedures
  • People know how to use the system
  • Train people on the culture of management
What are you trying to preserve?

- **Information** - The raw data or information contained in the record
- **Functionality/Appearance** - Look & Feel
- **Context** — How the records relate to other records
Information versus Appearance

2. Active Records

Records management can have the greatest positive impact when applied to active records. These are the records held by government because they are used most frequently, cost the most to maintain, and are essential to the current management of the government.

Activities Eligible for Support

a) **Files Management** supports projects to reorganize paper or electronic files and develop and implement files classification systems, to develop written policies and procedures, and to train staff. Eligible expenditures include file shelving (including that with locking covers), side-tab file folders and associated supplies. Fire-resistant filing cabinets are eligible only if their need is sufficiently justified, but filing cabinets (storage devices with drawers) and top-tab file folders are not eligible.

b) **Disaster and Business Recovery Planning** supports projects to develop and test disaster and business recovery plans.

(continued...)

Electronic Records Systems covers the development of needs assessment or the implementation of any record-keeping system not covered under another category. Such record-keeping systems could include database management systems (such as fire incident reporting software), computer output to laser disk (COLD) applications, and many others.

b) **Business Process Analysis (BPA)** supports the analysis and improvement of business processes that create or maintain records. BPA projects are a good choice for a government that has identified a problem with the way it conducts a specific record-keeping activity but does not have a specific technological solution to that problem.
The value of context

This is a GIS shapefile – one layer out of GIS system. By itself it’s just a bunch of dots. (Stockyards in KY)

All of the dots have data attached to them.

But without more map layers behind it, it’s just a bunch of dots.
Long-Term Preservation Strategies: Convert to Hardcopy (Eye-readable)

• Solution when
  – All necessary metadata is captured
  – No need to maintain functionality of records
  – Frequency of use goes down as time goes on
  – Color is not a major concern

• Examples
  – Print electronic data to paper
  – Digital to Microfilm

Best for text-based data:
Text files
Imaged paper
Digital Photos (mainly B&W)
Long-Term Preservation Strategies: Standard (Sustainable) Formats

Formats expected not to change, or change slowly

- Widely supported & used
- Easily Transferable
  - Compatible with other applications
  - Forward/Backward Compatibility
  - Version Control
- Non-proprietary or open/published standards

Types of Standards
- State/Enterprise
- National/International Standards (ANSI, ISO)
- Industry Standards
- “De facto” Standards
Types of Image formats

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>FILE EXTENSION</th>
<th>TYPE OF COMPRESSION</th>
<th>METHODS</th>
<th>USAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMP (bitmap)</td>
<td>.bmp</td>
<td>Cosiderably compressed with lossless</td>
<td>ZIP</td>
<td>used to store bitmap digital images</td>
</tr>
<tr>
<td>JPEG (Joint Photographic Experts Group)</td>
<td>.jpg, .jpeg, .jpe</td>
<td>Lossy</td>
<td>- Discrete Cosine Transform (DCT) &amp; Chroma Subsampling - Run-Length Encoding (RLE)</td>
<td>For natural images</td>
</tr>
<tr>
<td>GIF (Graphics Interchange Format)</td>
<td>.gif, .giff, .gfa</td>
<td>Lossless</td>
<td>LZW (Lempel-Ziv-Welch)</td>
<td>For artificial images (sharp-edge lines and few colors) &amp; support animation</td>
</tr>
<tr>
<td>PNG (Portable Network Graphics)</td>
<td>.png</td>
<td>Lossless</td>
<td>DEFLATE</td>
<td>Better compression &amp; features than GIF, but don't support animation</td>
</tr>
<tr>
<td>TIFF (Tagged Image File Format)</td>
<td>.tiff, .tif</td>
<td>Lossless</td>
<td>RLE / LZW / DEFLATE / ZIP</td>
<td>Flexible file format, can store multiple images in a single file</td>
</tr>
<tr>
<td>JPEG2000</td>
<td>jp2, .j2c, jpc, j2k, jpx</td>
<td>Lossy &amp; Lossless</td>
<td>Discrete Wavelet Transform (DWT)</td>
<td>Better image quality than JPEG (up to 20%), not widely used because of some patent issues.</td>
</tr>
</tbody>
</table>
Word about compression

• Digital media files come in 3 types:
  – Uncompressed – “raw” files in actual size
    • Audio – Wave (.wav)
    • Image - Tiff (.tff)
  – Lossless – Compressed but without loss of quality
    • Audio – Windows Media Audio (.wma)
    • Image – PNG (.png)
  – Lossy – Compressed but with loss of quality
    • Audio – Mp3
    • Image – JPEG (.jpg)
Cost

• **Storage**
  
  • IT Support and Maintenance
    – Unending and on-going
  
  • Backup/Replication Practices
  
  • Preservation Original
    – Distribution Copies
    – Preservation Versions

• **Data Volume**
  
  – Continuous increase without purging

• **Staff Resources**
  
  – Description, preservation, access
  
  – Backlog size
Volume - We can’t keep everything...

- Volume of records in electronic format is growing
  - Digital audio/video,
  - High resolution photos,
  - Databases

- Keeping everything diminishes the value of what is kept...
  - Strain on resources
  - Takes resources away from other projects
Final “official” version
176 GB

Raw or unedited “original”
2.21 TB

uploaded flash files
15.5 GB
Photos from Gov. Perry, Office of the First Lady

What do we know:
- Originals:
  - Raw, uncompressed camera file.
  - Proprietary format
  - All the pics taken
- Other 3 folders
  - All JPEG files of various size (compressed)
  - Only a sample of the total photos
  - But shows the ones actually used
Collaborating with IT

- IT maintains the technology
  - IT is the *physical* custodian
- IT does not own the data
  - Creator is the *legal* custodian
- IT focuses on maintaining the storage devices
  - IT does not consider long-term access/preservation
- Data owner must *define* its need for retention and *communicate* with IT
  - Data owners often *assume* IT is already addressing this need
“ARCHIVE”

THAT WORD

I do not think it means what you think it means
Building the Texas Digital Archive
TSLAC’s Mission

• Preserve the record of government for public scrutiny,

• Secure and make accessible historically significant records and other valuable resources, both for print and electronic documents,

Been having issues with the electronic part. Now looking to change that.
Enter Gov. Perry
Preservation System Functional Requirements

• Ingest electronic records from state agencies
• Characterize, validate and track digital formats over time
• Monitor integrity and security of digital files
• Migrate to sustainable formats for preservation
• Collect and store all appropriate metadata about the records
• Allow for proper access to records that meets the needs of our users
  – Provide records in access formats based on user needs
• Follow appropriate archival standards for preservation
Ingest Workflow – how we bring a file into the repository.

- Runs a series of “micro-services” on the files
- Most important for long-term preservation being the “Characterize” step.
• Characterize step identifies the file format
• Plugged in to an international registry of file formats
These are MS Word documents marked as the Original
Migration example

After migration now have a “Presentation” format in PDF.
https://www.tsl.texas.gov/texasdigitalarchive
The Texas Digital Archive (TDA) manages, preserves, and facilitates access to the electronic records collections of the Texas State Library and Archives Commission, including those transferred by State agencies or digitized by the State Archives. All records visible in this portal are unrestricted and available for public use.

**Locating records in the TDA**

- Use the links to the descriptive guides under "Collections" below to locate records in the TDA.
- Browse the TDA holdings by using the "explore archive" button at the top of the page.
- Searching is available through the "search archive" bar at the top of each page.
- For more information and help with using the TDA, go to the "search tips" page.

**Collections in the TDA**

For a listing of online collections not held in the Texas Digital Archive, please click here. For more information concerning our holdings, including materials not available online, visit the Archives & Reference page.

To return to the TSLAC main website, please click here.

When using maps and other records from the Texas Digital Archive for publication or display, please use the credit "Courtesy of Texas State Library and Archives Commission (and cite the collection name and record title or image number)."
TDA
Next Steps
Expanding our capacity for preservation

- Forensic Workstation to start working with legacy media
  - Hard drives
  - Floppy disks
  - Old formats
Building Partnerships with other state agencies

• Working with 3 new agencies:
  – Historical Commission
  – Office of the Attorney General
  – TX Parks and Wildlife

• Refining our transfer procedures
  – Contact us first
  – Fill out survey/inventory
  – Negotiate method(s) of transfer
What we are looking for – the short form

• Total volume (MB or GB) of “A” or “R” series
• What % of your volume is:
  – Databases
  – Audio/Video
  – Images
  – Unstructured data (Word,Excel,PDF, etc.)
• How are you storing your records?
  – Servers
  – Local network
  – Cloud
  – Removable media
  – Tape
  – Other
• What is the volume of the most valuable series that could be sent to the State Archives?
• How is the data being backed up?
• What issues are you encountering providing access to your electronic records?
• Would you be interested in transferring your archival electronic records to TSLAC?
Providing advice and assistance

• Assist agencies in managing their long-term records
  – Guidance documents & Best practices
  – Work with SLRM to develop training packages for long-term digital preservation
  – Serve as a knowledge base for state and local agencies
Questions?

Thank You!

Mark Myers
Senior Electronic Records Specialist
E-mail:
mmyers@tsl.texas.gov
https://www.tsl.texas.gov/texasdigitalarchive

Current Lobby Exhibit at TSLAC:
Greetings from Texas: A History of Postcards from the Lone Star State