


Creating Infrastructures For Long Term Digital Preservation

For Libraries, Museums
and Memory Institutions



Ray Uzwyshyn, Ph.D. MLIS
Director Collections and Digital Services
Texas State University Libraries
For Internet Librarian: Connect 2021



What is Long Term Library Digital Preservation Storage?

- Simply put, it is Very long-term Digital Storage, 10 years or more.
- Libraries, Special Collections and University Archives collect print materials & other media. Increasingly, they gather and collect digital information, media and data.
- There are new millennia archival perspectives and digital storage standards in line with Research library standards (ISO standards: **16363**, **16919**, **14721**)

Digital Preservation in Libraries follows a Unique 3-Legged Stool Library Model

Organization

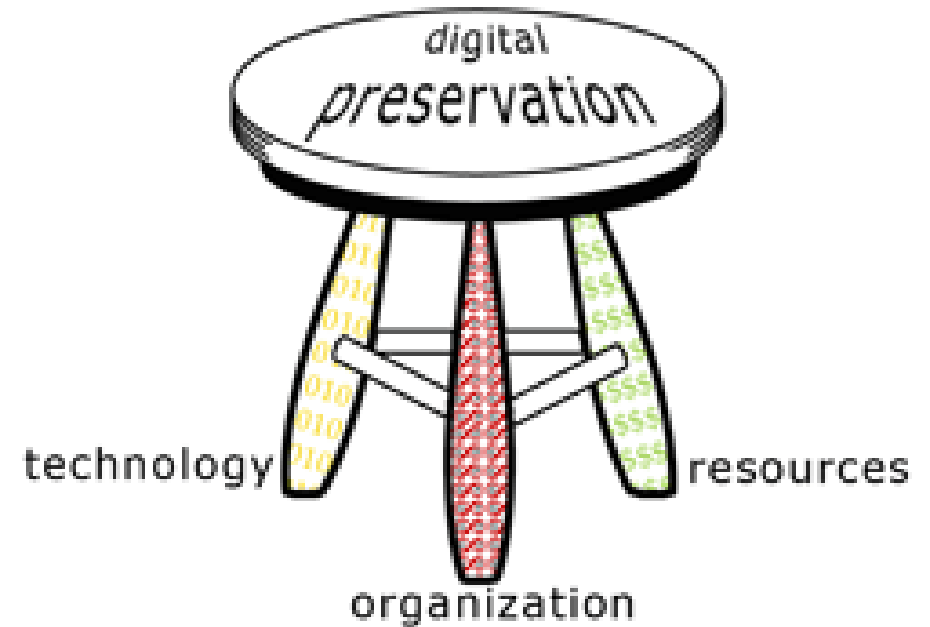
Leverages existing human and consortial resources in libraries to build on archival/stewardship expertise

Technology

Synthesizes Technological Capabilities to meld with Traditional Library Archival/Collection Preservation Models

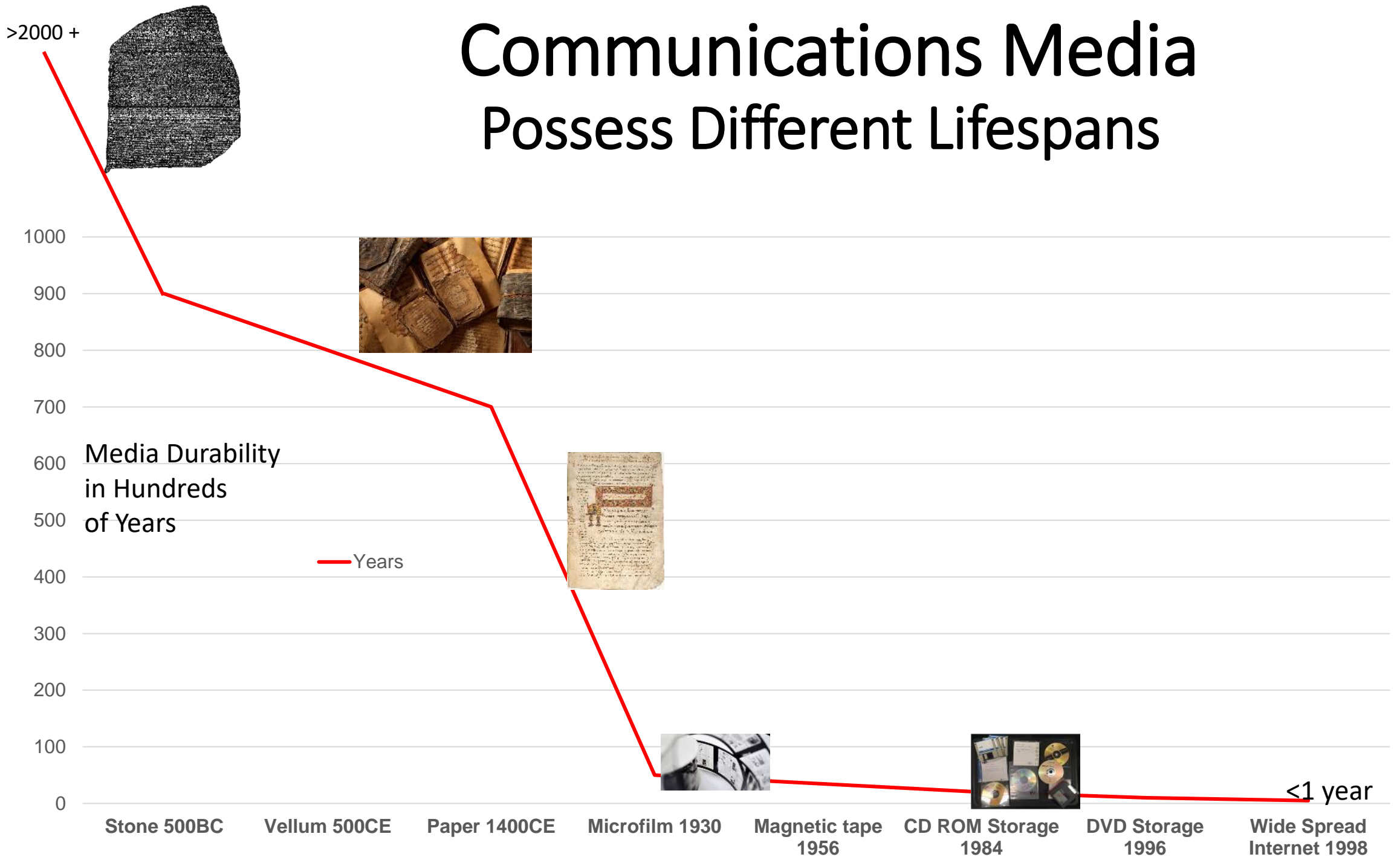
Resources

Utilizes Library, IT, Network and Human Resources.



[Anne Kenney/Nancy McGovern, 2007](#)

Communications Media Possess Different Lifespans



Many Considerations For Long-Term Digital Preservation Solutions

- Technological
- Disaster Planning
- Institutional Failure



Any Solution Must Allow For

Technological Diversity of Formats

Digital Replication (LOCKSS Principle)

Digital Auditing and Repair (Bitsums)

Geographic Distribution

Best Practice Standards

Succession Agreements

Unique Characteristics of Technology of Long-Term Digital Preservation



Any Digital Preservation Technology Must Allow for:

- Migration and Preservation of Formats for Long Term Storage (Normalization of Files, Migration Forward)
- Risk Mitigation for Data and Content.
Multiple bit-level copies, stored in disparate locations geographically, administratively, and technologically.
- Leveraging the libraries' role as keeper of the scholarly and historical record now in a digital arena

Primary Steps

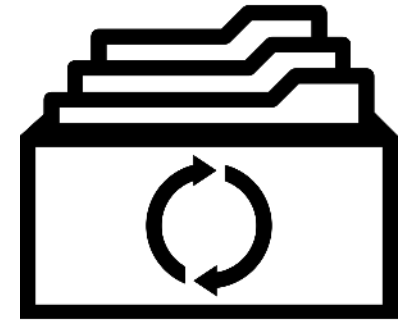
Step 1: Form a Digital Preservation Working Group
(Texas State University Libraries Example)
TXU DPWG Background & History

Purpose:

A Digital Preservation Working Group provides oversight, direction and responsibility for:

- 1) Digital Preservation Technology
- 2) Policy Infrastructure

- **Group Formed 2015** and consists of members of Libraries' Digital and Web Services (Digitalization Lab, Institutional Repositories) University Archives, Wittliff Special Collections, Library General Collections
- Group began by investigating and then authoring the Libraries' first [Digital Preservation Policy Document](#) (August 2016), benchmark minimums for preservation Masters etc.
- Created Dedicated Local Server Space for Preservation Files and Use Files with Library IT/University Technology Resources



Technology



Policy

Investigate New Digital Preservation Tools, Platforms and Resources

- **Archivematica:** Middleware standard for Digital Preservation Metadata and Integrity
 - Archivematica bundles micro-services for normalizing files, managing metadata and verifying file types, bit-level integrity (checksums) etc.
 - Texas State Began R&D with Archivematica on Linux Ubuntu and first deployed production level instance on a Linux Red Hat platform
 - University Archives and Special Collections began experimenting with, learning and utilizing Software (2016-2018)
 - All areas gained expertise in Metadata, middleware workflow process (Archivematica) to create AIP's (Archival Information Packages) to safely store, archive and retrieve files and metadata for later use

@archivematica



Step 2: Conduct Initial Digital Storage Needs Estimate (DPWG)

- **University Archives:**

- Thesis & Diss. project: 500 GB per year
- Yearbook/Football negatives: 235GB per year
- San Marcos Daily Record Negatives 1500 GB per year
- Audio digitization: 500 GB per year.
- Misc imaging: 500GB per year



- **Special Collections (Wittliff):**

- Unique digitization projects. Lonesome Dove Dailies (20 TB), Powers (10 TB) , Broyles (300 GB). Jerry Jeff Walker 2# reel tapes .
- O'Connor Collection/New Major Donation example (2TB).
- Austin Film Festival: 1.5 TB per year, (2+ years).
- Misc imaging: 2 TB per year
- Audio digitization: Wittliff: 200 GB / year



- **General Collections:**

- Theses and Dissertations
- Streaming media archive: 2 TB per year, General Collections (Not Covered by LOCKSS, PORTICO Memberships)

Conclusions ~ 60-70 TB Initial Digital Storage, required
10-12 new TB/year for access files needed



Step 3: If Possible, Join Consortia

Texas Digital Library

Forms First State Digital Preservation
Resource Infrastructure (2016-2018)

- **2016 TDL Preservation Services Initiated**
(Hires Courtney Mumma from Internet Archive (Wayback Machine, Brewster Kale) to Focus on State [Digital Preservation Services](#))
- **2016 TDL Forms Alliance with DuraCloud**
(Digital Preservation focused Non-Profit [Duracloud @ TDL](#))
- **2017 TDL Creates Digital Preservation Services**
Members receive “[Space](#)” in DuraCloud@TDL for ingesting content, based on [membership level](#).
- **2018-2019 Texas wide TDL Digital Preservation Groups and State Committees Formed**
[Archivematica Users Group](#) Formed
- **2019 TDL State Digital Preservation Committee Formed**



Step 4: Storage Infrastructure Recommendation

2018-2019 Digital Preservation Working Group
Continually Changing Landscape

The Best
**Unlimited Cloud
Storage Solutions**



Recommendation Charge 4 Pillar Methodology

- 1) Conduct Environmental Scan:** to Identify Library Digital Preservation Storage Options
- 2) Compare Peer Groups (TDL)** and National Best Practices for Research Libraries
- 3) Narrow Focus** to Pragmatic options suitable for University Libraries Needs
- 4) Forward Recommendation:** for AVP and VPIT Review and Approval



Pillar 1: Environmental Scan

Digital Preservation Solutions (Peer Institutions)

Texas Peer Institutions	University of Texas at San Antonio	University of Houston	UT Rio Grande Valley	University of Texas (Austin)	Texas A & M University
Digital Preservation Solutions	Duracloud Directly (not via Texas Digital Library, TDL)	Amazon S3 and Glacier Directly (Not via Texas Digital Library, TDL)	Chronopolis via DuraCloud through TDL	LTO Tape, moving to Texas Advanced Computing Center	Chronopolis and Amazon via Duracloud @ TDL

Pillar 2: Narrow Focus

Three Final Candidates

Option 1: Outsource Preservation Digital Storage

- Preservica

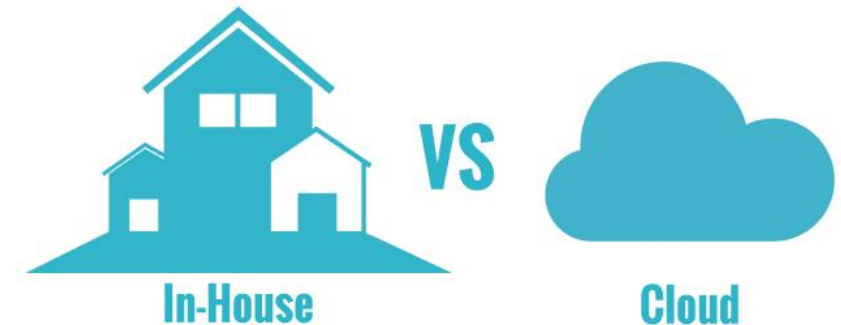
Option 2: In-House Texas State Data Center Solution

- files.txstate.edu

Option 3: Mixed Duracloud through TDL/Other

Options

- AmazonS3
- Amazon Glacier
- Chronopolis
- (Azure)



Option 1: Outsource

(All in One Outsource Option, Preservica)

Benefits	Considerations
Preservica creates AIP's (Archival Information Packages, Metadata) and provides all technology set-up and support	Costs: \$35,000.00/year for 20TB
Established Archival Best Practices	No local control or entrance to underlying technology (black box)
Recognized Library Peer and Community of Practice	Variable Response to Local Needs (similar considerations to @mire)



Option 2: In House

Expand TR/Texas State Data Center Relationship

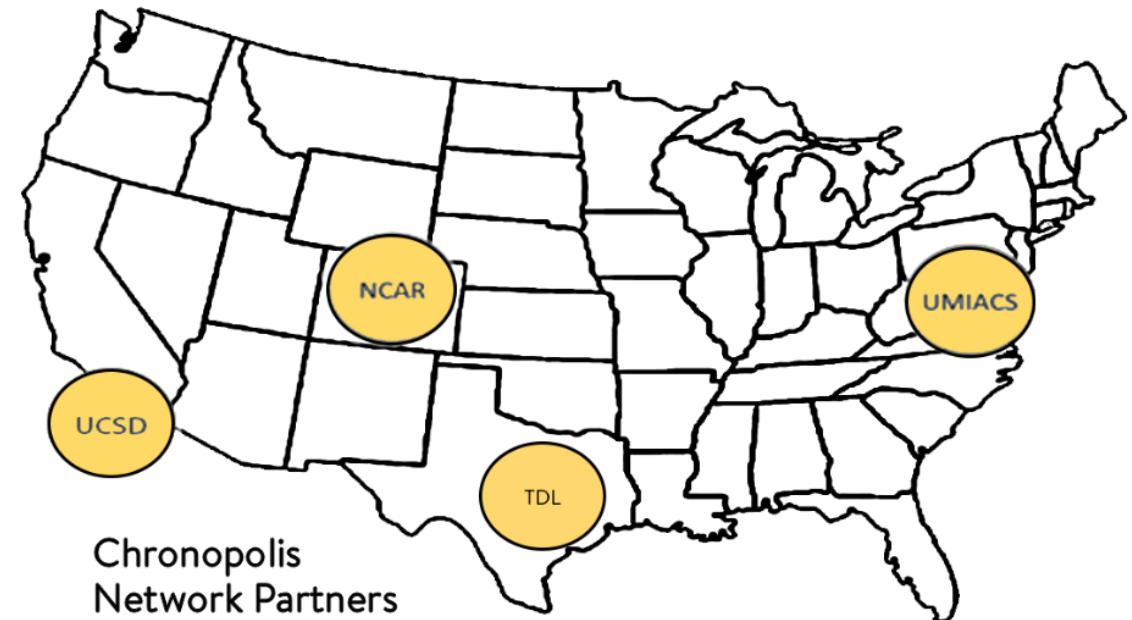
Benefits	Considerations
Proven relationship with University Technology Resources.	Specialization not in place: Metadata Infrastructure, Normalization of Formats, library-related expertise or best practices for this type of Digital Preservation
Storage for working files , access copies, preservation files and associated metadata established	Requirements for geographic, administrative and technological distribution (even if multiple copies) currently not met
Building on our current temporary solution of files.txstate.edu and increasing capacity. Growth estimate of 10-12 TB/year	30-day window for recovery is currently not sufficient for maintaining long term preservation files and associated infrastructures needed

Option 3 :Mixed Duracloud through Texas Digital Library to Chronopolis

- **Chronopolis: Geographically Distributed Preservation Network**
 - UC San Diego
 - National Center for Atmospheric Research
 - University of Maryland, Institute for Advanced Computing Studies
 - TACC (Texas Advanced Computing Center)



Benefits	Considerations
Geographic Distribution at any 3 technologically diverse partner nodes	Subscription cost: \$2500 annual fee includes 2TB/year storage and ingest \$1000 initial setup (<i>1st year only</i>)
Non-Commercial solution rooted in libraries and cultural heritage community	Storage \$165/year/additional TB \$120 ingest fee/additional TB
Library community of practice around this (TDL/Duracloud/Chronopolis)	Significant Human resources/time investment for initial technological integration
File Fixity and Data Integrity processes are transparent	

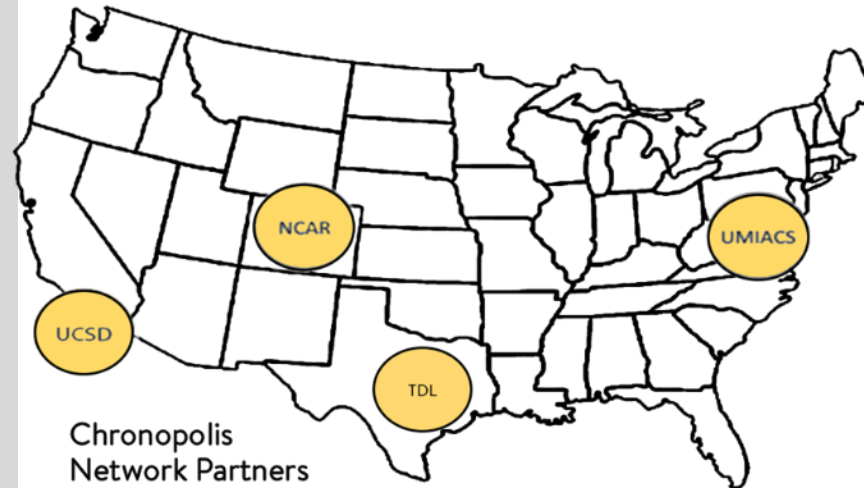
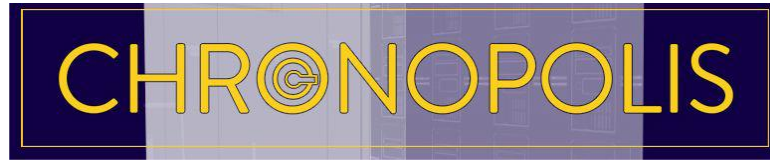


Option 3: Mixed

Duracloud through TDL (Texas Digital Library) to Chronopolis Option



- Provides strong library support through four academic library focused organizations (Chronopolis, TDL, Duraspace, Lyrasis) for long term viability and strong peer support networks
- Anticipated Budgetary Request:
 - Year 1: **\$3500.00** (\$2500.00 TDL Preservation/year, \$1000.00 Initial Set-up/Onboarding, Includes 2 TB Storage)
 - Year 2-3: **\$2785.00/year** (includes additional 1 TB storage/year)
- Review Storage and Staff Needs Annually.



Future Directions (2022 +)

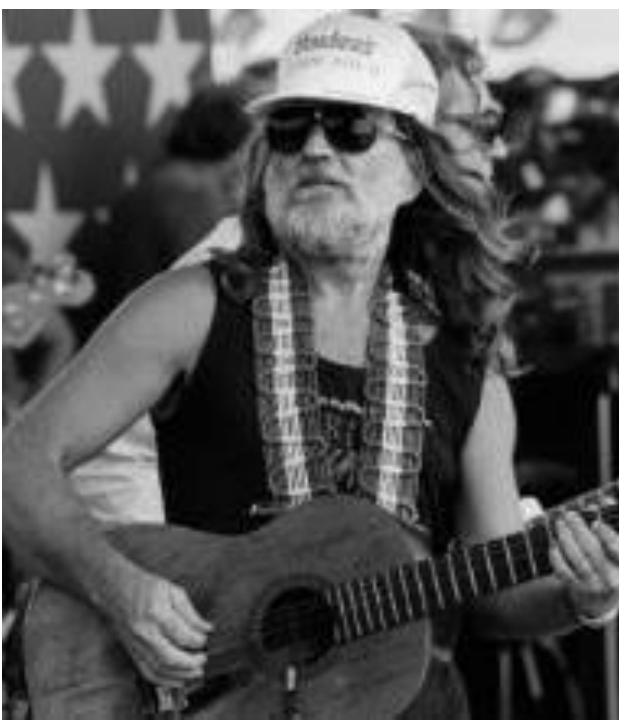
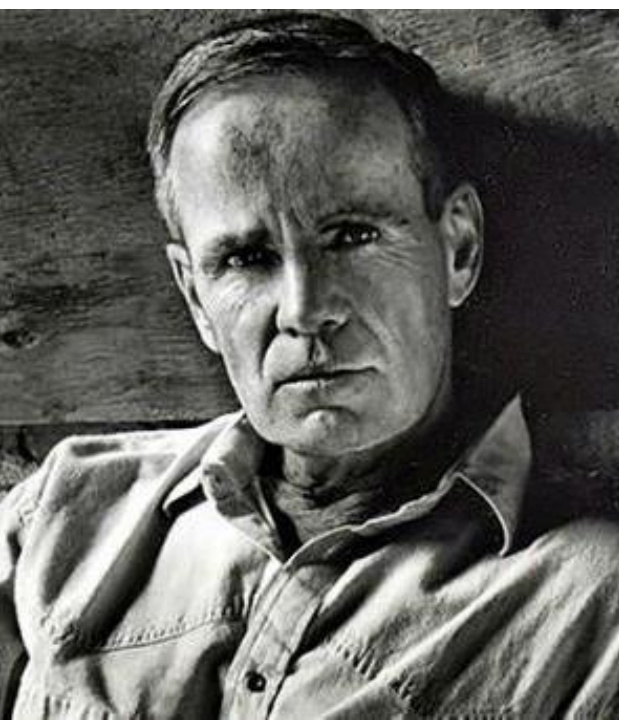
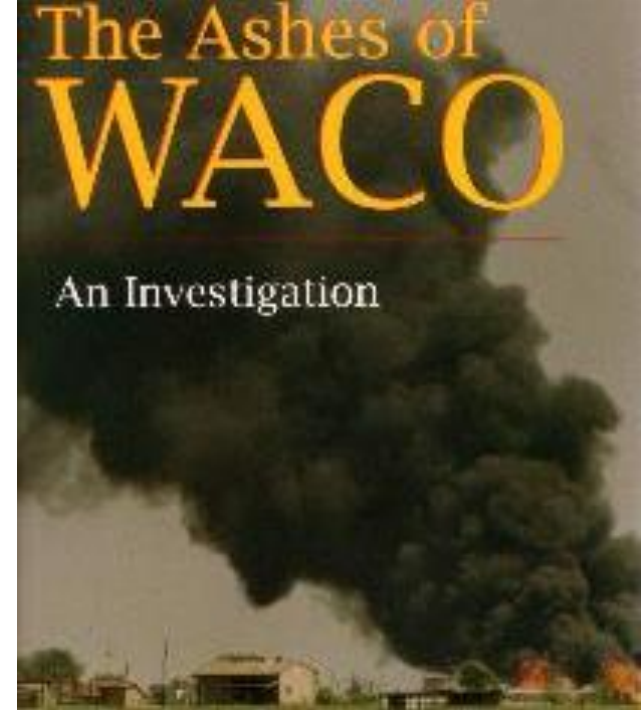
Web Archiving, Digital Forensics, Email Archiving

Web Archive Life Cycle



Web archiving = The process of collecting portions of the World Wide Web to ensure information is preserved in an archive for future researchers, historians, and the public.

- Employs web crawlers for automated capture.
- Largest web archiving organization based on a bulk crawling approach is the Wayback Machine (Internet Archive) <https://archive.org/web/>



Web Archiving Opportunities Archives and Special Collections 2022-2027

Wittliff Special Collection Four Pillar Priorities: Texas & Southwestern US

Literature, Film, Music Photography

- **Crawling quickly disappearing** review, article and dedicated fan sites, other areas of interest regarding collection priorities



University Archives

- **Crawling txstate.edu** for what websites/areas are of interest Partnering with Texas State University researchers, colleges, departments, institutional records
- **Potentially beginning** archiving Texas State important **Social Media** Twitter, Facebook

Web Archiving R&D Phase

Platforms – Archive-IT, WebRecorder

- Launched in 2006
- Built by the Internet Archive (Wayback Machine)
- End-to-end hosted platform to create, store, and provide access to collections of web content
- Most widely-used platform for universities, <https://archive-it.org/blog/learn-more/>
- Pricing
 - Different pricing levels based on amount of data archived annually
 - TDL is investigating consortial discounts as well as other platforms (WebRecorder)



WebRecorder

Conclusions and Deeper Rationale For Long Term Digital Preservation Storage Infrastructure

Long Term Digital Preservation

Provides a New Level of Service Expected by Donors, Researchers, Faculty and students.

Is a Necessary Focus Area for Research Libraries

Connects The Library with State & National Library Technology Organizations

(Texas Digital Library, Digital Preservation Network, Coalition of Network Information, Lyris, Chronopolis, Duraspace)

Places Your Library in Line with other leading-edge institutions working in these new areas

(Association of Research Libraries, ARL, GWLA, Greater Western Library Association, HathiTrust, CNI etc.)

Long Term Digital Preservation Software and Storage Related Links

Archivematica. <https://www.archivematica.org/en/>

Amazon Web Services Cloud Storage
<https://aws.amazon.com/products/storage/>

Chronopolis Digital Preservation Network.
<https://aws.amazon.com/products/storage/>

Duracloud Digital Preservation
<https://duraspace.org/duracloud/>

Microsoft Azure <https://azure.microsoft.com/en-us/>

Preservica <https://preservica.com/>

Texas Digital Library Digital Preservation Services:
<https://www.tdl.org/digital-preservation/>

Long Term Digital Preservation Standards, Resources, Articles and Presentations

ISO Policies. #16363 Audit and certification of trustworthy digital repositories. <https://www.iso.org/standard/56510.html> , #16919 Requirements for Trustworthy Digital Repositories. <https://www.iso.org/standard/57950.html> , #14721 Open archival information systems (OAIS). <https://www.iso.org/standard/57284.html>

Kenney, A and McGovern, N. A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation? RLG DigiNews April 15, 2007. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/60441/McGovern-Digital_Decade.html?sequence=4

Library of Congress. Digital Preservation at the Library of Congress. (Retrieved 2021) <https://www.loc.gov/preservation/digital/>

Texas State Digital Preservation Working Group. (Retrieved 2021) **Texas State University Libraries Digital Preservation Policy.** <https://www.thewittliffcollections.txstate.edu/research/visit/policies/dig-pres-policy.html>

Uzwysyn, R. (2021). Building Frameworks for Long Term Digital Preservation. **Computers in Libraries.** September, 2021. Vol. 41, Number 7. pp. 4-8. ISSN: 1041-7915. <https://www.infoday.com/cilmag/sep21/Uzwysyn--Building-Frameworks-for-Long-Term-Digital-Preservation.shtml>

Uzwysyn, R. (2020). Digital Preservation Storage Infrastructures Model Proposal Presentation. Texas State University. DOI: [10.13140/RG.2.2.14102.09289](https://doi.org/10.13140/RG.2.2.14102.09289), https://www.researchgate.net/publication/339390854_Long_Term_Digital_Preservation_Storage_Infrastructures_for_Libraries_Archives_and_Research_Institutions?channel=doi&linkId=5e4f04dd299bf1cdb9391aeb&showFulltext=true

Questions and Comments

Ray Uzwyshyn, Ph.D. MLIS MBA
email: R_U15@txstate.edu
<http://rayuzwyshyn.net>
Director Collections and Digital Services
Texas State University Libraries



Step 4

Digital Preservation Storage Recommendation Focus

- Investigation begins into various Historic, Library Centered, University and Commercial Storage Solutions
- DPWG Group gaining recognition, awareness of permanent digital preservation storage needs, capabilities of libraries
- Resource possibilities maturing and widely available commercially and in the library space
- Possible solutions ranged from new to historical models to In-House and Outsourcing possibilities

Digital Preservation Storage Working Group Final Recommendation



Option 3: Duracloud through TDL

Amazon S3 and Glacier Option

Benefits	Considerations
Amazon S3 suitable for streaming, dynamic access. Amazon Glacier suitable for long-term dark archive needs	Commercial: not tailored to cultural heritage institutions. Does not meet requirements for geographic, administrative and technological distribution
Amazon Glacier and Amazon S3 are both part and options within the Duracloud Suite if we ever chose to use them	File fixity and data integrity is a black box (process hidden from owners)
TDL and Duracloud both possess established community of library best practices.	Subscription cost \$2500 annual fee includes 2TB/year \$1000 initial setup (<i>1st year only</i>) S3 \$265/year per additional TB Glacier \$50 / year per additional TB
	HR/Time Investment for Initial Technological Integration



Option 3:

Duracloud Component

- **Duracloud** is a hosted middleware service from DuraSpace that lets organizations control where and how digital content is preserved.
- The parent organization **Duraspace** is a non-profit organization providing academic library leadership for open source technologies focused upon durable, persistent access to digital data. (i.e. Fedora, Dspace).
- Currently, Duraspace is part of **Lyrisis**, a longstanding library related organization supporting libraries and technology initiatives



Option 3: Duracloud Through the Texas Digital Library (TDL)

- Duracloud would be administered through our TDL membership with these consortial relationships, advantages (usergroups, networks etc.) and constraints
- The Texas Digital Library is a Consortial Organization consisting of 22 Texas University Library Organizations
- Focused on enabling Texas Libraries Digital Infrastructure and new digital technology Projects.



Conclusions

and Deeper
Rationale
For Long Term Digital
Preservation
Storage Infrastructure

Long Term Digital Preservation Provides a New Level of Service Expected by Donors, Researchers, Faculty and students.

Long Term Digital Preservation is a Necessary Focus Area for Research Libraries

Long Term Digital Preservation Connects Library with State & National Library Technology Organizations
(Texas Digital Library, Digital Preservation Network, Coalition of Network Information, Lyris, Chronopolis, Duraspace)

Places Your Library in Line with other leading-edge institutions working in these new areas
(Association of Research Libraries, ARL, GWLA, Greater Western Library Association, HathiTrust, CNI etc.)

