Building Frameworks for Long Term Digital Preservation

Texas State University Libraries, Alkek Buck Wynn Historical Mural and Information Commons, https://www.library.txstate.edu/

By Ray Uzwyshyn, Ph.D. MBA MLIS
Director, Collections and Digital Services, Texas State University Libraries

Libraries are memory institutions and have unique historical roles as stewards in the preservation of our collective histories and knowledge production. Building frameworks for long term digital preservation storage infrastructures in libraries involves groups, institutions, IT departments, librarians and archivists. In our new millennium long term digital preservation infrastructures are key areas for memory institutions. Information comes up online and goes down offline and then disappears, sometimes forever. Future proofing requires longer term thinking, not just a one-year project management plan but longer-term foundations, scaffolding and thinking. This article offers such a foundation and pragmatic structured framework for putting together solid long term digital preservation infrastructures.

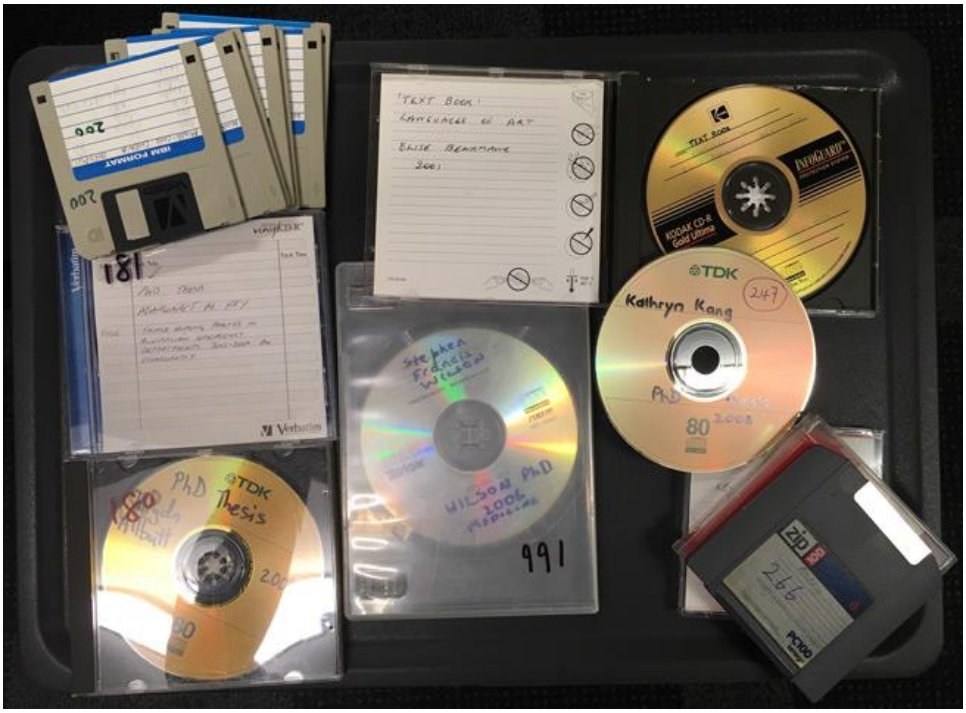
Frameworks for Digital Preservation

Options for long term digital preservation such as outsourcing, working in house, staff hires, and hybrid software combinations can be overwhelming. What models and best practices have libraries developed for long term digital preservation? What exactly are long term digital preservation frameworks? This article defines and then focuses on a best practice framework for digital preservation

currently implemented at Texas State University Libraries through its Digital Preservation Working Group. Whether you have plans to put together this type of system or have already assembled a preliminary digital preservation system, this overview advocates a structured approach. We will begin by briefly examining what exactly long-term digital preservation is and how to conduct a local digital storage needs estimate. Processes involved in conducting an environmental scan to make a storage provider recommendation will be reviewed. Finally, deeper rationale for long term digital preservation storage will be weighed in terms of present needs and long-term considerations.

What is Long Term Digital Preservation Storage?

University Libraries, Special Collections and archives increasingly collect and gather information in both analogue and digital formats. Well-known analogue information ranges from books, paper-based archives to videotapes and LP's. Digital information ranges from digital PC files, DV media, data and increasingly, websites and email. Much of this information could benefit from digitization if it is in an analogue format and longer-term digital storage and retrieval if it is already in a digital format. Digital preservation storage for libraries is 'very' long-term digital storage beyond standard IT department disaster recovery and regular 'records retention' mandates. It is a commitment to a larger and longer program that involves both human resources and budgetary allocations. Complex ISO standards define needs for the audit and certification of trusted digital repositories and open archival digital information storage systems (ISO #'s 16363,16919,14721, See references) but currently there are only more than a handful of trusted repositories that completely meet these high benchmarking systems' requirements.



Digital Media Files consist of a wide variety of file formats, types and standards.

3-Legged Stool Digital Preservation Model.

Pragmatically, most libraries pursuing digital preservation follow a 3-legged stool model originally suggested by Anne Kenney and Nancy McGovern in 2007. A solid foundation for digital preservation is built upon a three-legged stool consisting of technology, organization and resources. **Organization** leverages the respective organization's existing human resources to build on their archival/stewardship expertise for the digital age. **Technology** synthesizes current technological capabilities with traditional library archival/collection preservation models. **Resources** seeks to use library human resources and wider network resources like IT departments or national alliances and consortia possibilities.

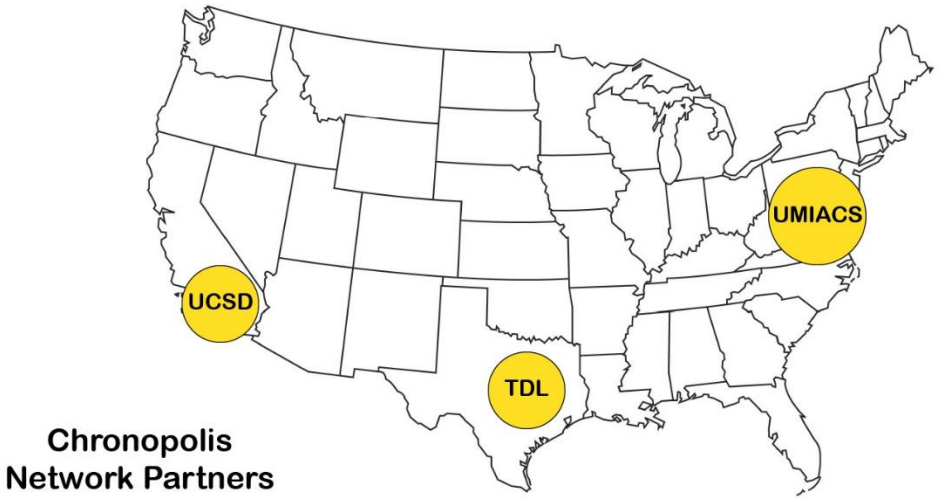
Digital Preservations Three-Legged Stool Model (See Anne Kenney/Nancy McGovern, 2007)

The Unique Characteristics of Long-Term Digital Preservation

The unique characteristics of long-term digital preservation hinge on the migration and preservation of digital file formats for long term storage. This means that a Windows 3 Word 2 file from 1991 can be read as easily as an Office 360 Word file from 2021. This ability to 'make readable' is called 'normalization', migrating file formats forward to be read normally within 'present day standard' technological formats. The work of preservation and normalization is an ongoing process, continually moving file formats forward to present 'readability' standards.

Migration and Normalization of File Formats for Long Term Preservation. Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Risk mitigation for data or content comes as a close second unique characteristics of long-term digital preservation. Pragmatically, risk mitigation means creating multiple bit-level copies of a file and doublechecking for bit-level accuracy (technically called checksums) and then storing these files in disparate locations geographically, administratively and technologically.



Chronopolis is a geographically distributed preservation network among University of California San Diego (UCSD), Texas Digital Library @ Texas Advanced Computing Center (TDL@TACC) and University of Maryland Institute for Advanced Computing Studies (UMIACS), https://library.ucsd.edu/chronopolis/index.html

Building a Long-Term Digital Preservation Working Group

For institutions thinking about long term preservation infrastructures, it is best to begin by forming a digital preservation working group. This group will consist of members of the libraries and IT departments digital and web services (I.e., digitization lab, institutional repositories, core IT,) and the University libraries archives, special and general collections (archivists, collection development librarians and administrators). The group's tasks should begin by investigating and authoring a local digital preservation policy document. This document will clarify and overview an institution's Digital Policy Framework (purposes, objectives, roles and responsibilities), outline general procedures (digital preservation strategies, technological infrastructure) and provide a detailed plan, strategy and timeline for implementing and publicizing the framework. Fulfilling tasks of outlining policies and agreeing to benchmark minimums for preservation masters will also establish common ground for later more contentious group discussions. A model for such a plan is Texas State Universities Digital and Preservation Policy Document available here: https://www.thewittliffcollections.txstate.edu/research/visit/policies/dig-pres-policy.html

Texas State Digital Preservation Working Group Policy:
https://www.thewittliffcollections.txstate.edu/research/visit/policies/dig-pres-policy.html

Digital Preservation Tools

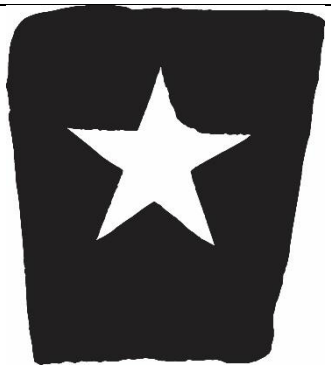
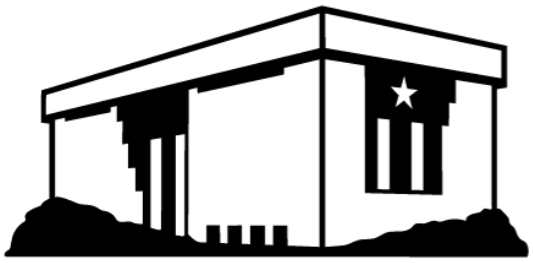
Next on any agenda for any Digital Preservation Working Group will be an environmental scan of the current landscape of digital preservation tools to decide on the right tools. In its own scan, the Texas State Digital Preservation Committee decided upon the open-source tool, Archivematica, a middleware software for ensuring file integrity and digital preservation metadata (See references, Uzwyshyn (2020)). Archivematica bundles micro services for normalizing files, managing metadata and verifying file types for bit-level integrity (i.e., the checksum). The Texas State Group first deployed production level instances on the Archivematica Red Hat Platform, a widely used open-source Linux platform. After implementation, archives and special collections staff may begin experimenting to gain expertise in this middleware's workflow process to create AIP's (Archival Information Packages) which store, archive and retrieve files and metadata for later use.

Archivematica, digital preservation middleware software: https://www.archivematica.org/en/

Developing a Digital Storage Needs Estimate

Simultaneous to developing the above digital preservation policy documents and software expertise, a Digital Preservation working group will also need to conduct an initial storage needs estimate. This storage estimate will both estimate the current permanent digital storage needed and the yearly additions estimated through current digital preservation work/year. It is good to develop this estimate early on to give to University IT and/or Library administration. This will involve cost estimates for data

storage and ongoing effects for the annual operating budget. It is also beneficial to give a high-level breakdown of current storage/space requirements by areas (I.e., Archives, Special and General Collections). An example of such a table is given below.

Texas State University Digital Preservation Group Digital Storage Needs Estimate **Conclusions:** 10-12 TB/year recurring and 60-70TB permanent Digital Storage, needed	
	**University Archives:** Thesis and Dissertations project: 500 GB per year Yearbook/Football negatives: 235GB per year San Marcos Daily Record Negatives 1500 GB per year Historical Audio digitization: 500 GB per year. Misc imaging: 500GB per year
	**Wittliff Special Collections:** Austin Film Festival: 1.5 TB per year, (2+ years). Lonesome Dove Digital Video Miniseries Dailies (20 TB), Jerry Jeff Walker 2# audio master reel tapes. Audio digitization (Selena)f: 200 GB / year Powers (10 TB), Broyles (300 GB). O'Connor Collection/New Major Donation example (2TB). Major Author Archives: 2 TB per year, Cormac McCarthy, Sandra Cisneros
	**General Collections:** Streaming media archive: 2 TB per year, General Collections Digital Serial Backfiles (Storage space needed but not covered by LOCKSS, PORTICO, Hathi Trust Memberships)

Digital Preservation Storage Solution Recommendation

With preliminary foundations and general needs established, it is also up to the Digital Preservation group to review the current landscape of storage solutions to forward a recommendation to library administrators and IT. This work involves conducting an environmental scan to identify library digital preservation storage solutions and conduct a cost benefit analysis. Cost per Terabyte or Petabyte per year will be particularly crucial factors as any solution technically entails digital preservation storage and retrieval in perpetuity and this will be an ongoing cost. It will also be beneficial to begin by looking at

both in house solutions and comparing those to cloud possibilities (Amazon Web Services, Azure, Chronopolis, Duracloud etc.). It will also not be a bad idea for the Preservation group to compare a peer group of libraries or other memory institutions pursuing digital preservation for national best practices and current trends. From here the group may narrow its focus to pragmatic options taking in different variables to suit the specific library or institution's needs. A presentation should be developed at this stage surveying the groups progress (i.e., policy and software, expertise developed) and from here a recommendation forwarded to administrators that will review the presentation proposal for future budgetary funding allocations for this new infrastructure. Our Texas State group's process and final choice led to Chronopolis via Duracloud through the Texas Digital Library. (A detailed presentation of this decision-making process is available in the reference links, See Uzwyshyn 2020). Together, following these steps outlined above will lead to a good initial foundation for any digital preservation framework.

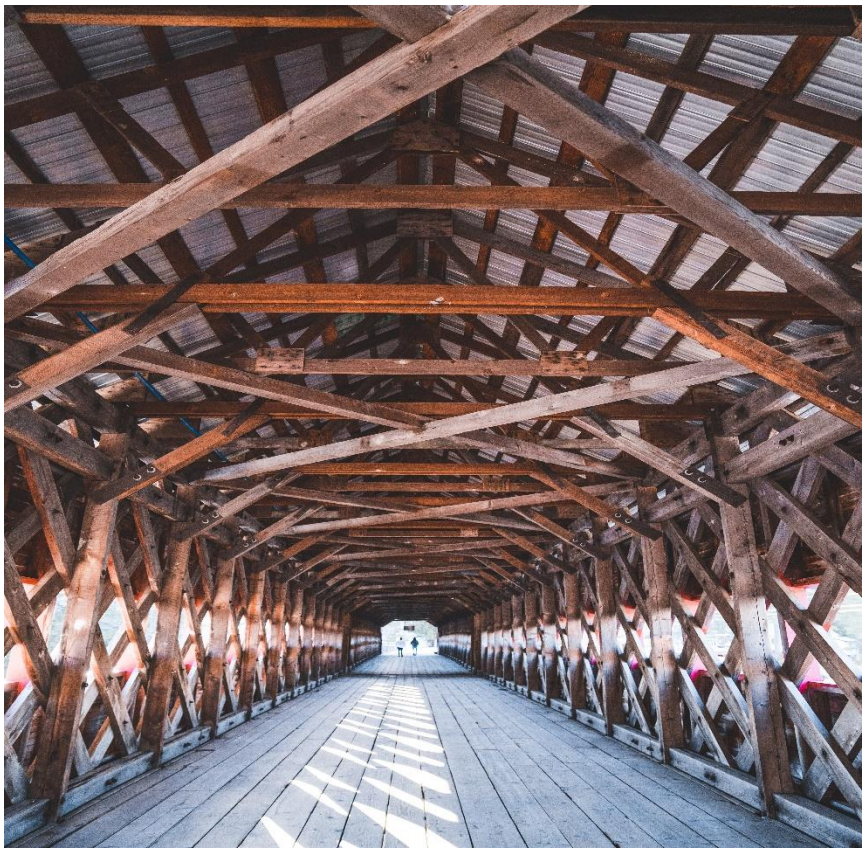
Duracloud, open-source technology project for preserving and archiving digital content: https://duraspace.org/duracloud/ and Texas Digital Library Digital Preservation Services: https://www.tdl.org/digital-preservation/

Conclusions

A new level of service is currently expected by researchers, university faculty, students and donors in our networked age. Digital preservation is one of the grand necessary challenges of our times. These new processes will bring relevance and a raison d'etre to any memory institution's existence in our digital era. By placing one's library into this arena, we also reconnect our libraries to longer threads of past historical legacies and lineages. In this way we may connect our present to institutions who have defined our past histories through their longer-term preservation decisions and define our future for which we lend on our voices and acts in shaping today.

Currently in the West, we are already 40 years into the era of digital preservation. This era began with the advent of the desktop computer as a way of working in the early 80's. The multitude of files and

important historical digital information produced since then now needs digital preservation. A wide spectrum of media formats is obsolescent or rapidly aging. These needs also continue into our present web centric world with the web and cloud's ever-burgeoning long term digital preservation needs. It is important for libraries and memory institution to begin wading into these waters. Information is disappearing daily. Important histories are being lost for our collective future. It is also important to remember too that Rome was not built in a day and a journey of a thousand miles begins with a single step. We are in a digital renaissance or Gutenberg II-like period. Historical digital information artifacts are at a stage akin to that of the early Incunabula in the first 50 years after the invention of these earlier communication model paradigm shifts. Longer term digital preservation as these early digital artefacts begin to disappear is needed now.



Frameworks Photo Courtesy of **Jean-Daniel Francoeur**, **Pexels**

Digital Preservation Standards, Resources 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

Uzwyshyn, R. (2020). Digital Preservation Storage Infrastructures Model Proposal Presentation. Texas State University. DOI: 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

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/