Challenges in the Preservation and Archiving of Digital Materials

Leslie Johnston
National Archives and Records Administration (NARA)
Community Webs Meeting, June 21, 2022

What are the key challenges for the archiving and preservation of born-digital research and scholarship, digitization products, and digital collections in general?

Heterogeneity

- Work and research within and across organizations utilize very different methodologies, equipment, software, and hardware. Outputs range from publications to websites, A/V, textual and numeric datasets, and software needed to process the results. This also applies to electronic records (born-digital or digitized) and general digital collections.
- Personal "papers" are increasingly born-digital and exist only in online platforms such as Facebook, Twitter, Instagram, YouTube, Gmail, iCloud, Dropbox, etc., most likely in all of them.
- There are literally thousands of variant versions of file formats over time, and they just keep changing. And we cannot identify every legacy format with certainty.
- There are dozens of current and legacy carrier formats—floppy disks, hard drives, CDs, DVDs, thumb drives, tapes, etc.—and we need to be able to read the files off them to preserve them.

Technology

- With heterogeneity comes a wide variety of ever-changing tools and workflows needed to process, describe, preserve, and provide access to born-digital scholarly research.
- Storage is much less expensive than it used to be — especially in the cloud— but it can become a budgeting concern when you consider scale and the need for preservation replication. And with cloud use comes egress costs for copying/moving files out.
- With scale also comes stress on local networks and the limiters of moving files using web protocols.
- Machines used to process born-digital materials will require increasingly more storage and memory and higher bandwidth network connections.

Complexity

- Digital materials do not exist without a context and a provenance which must be recorded and maintained.
- This is especially true for web sites, where the context and relationships between the hundreds or thousands of files that make up a site are vital for access.
- Scholarly output and electronic records are increasingly complex, comprised of multiple or multi-part or containerized files that require all their components, have relationships to other files, or are bundled with software that is necessary for research to be reusable and replicable.

Scale

- There are thousands of faculty, students, and prominent individuals associated with any university, organization, or community whose files will be collected by cultural heritage institutions over time.
- There is a massive amount of observational data and research datasets created in scientific research that research data preservation policies require, that the organizations researchers are affiliated with, must potentially retain and preserve.
- Some types of collections – audio, video, film, email – produce both huge files and huge numbers of files to preserve.

Access

- How do you describe and provide access to so many different types of objects and records?
- What do you provide to researchers from your catalogs?
- What are the “right” file formats?
- How much technical assistance can you provide?

Serving Multiple Communities and Purposes, Including Ourselves

- If it's not accessible, we have not preserved it.
- It's not just about the files and the technology, it's about people. There is no single community of creators, nor of users. And new communities will emerge.
- As with all our collections, we will never know all the uses that our digital files will serve for research or the public.
- We will need to change our own organizations to meet the needs of our collections and our communities.

It’s not all gloom and doom.

What are some of the successful strategies that should be part of every Digital Preservation program?

Guidance for Content Creators

- The digital preservation life cycle starts with the people creating the files, not when the files come over the transom to our organizations.
- There is no such thing as the ability to completely enforce what is created or what is collected, because the work requires whatever the appropriate tools or formats are. But guidance on data management strategies, appropriate storage criteria, preferred and acceptable file formats, and minimum metadata make long-term preservation more likely.
- Examples include Research Data Plans, Format Statements, FADGI Guidance, the NDSA Preservation Storage Criteria and Levels of Digital Preservation, etc.

Ongoing Risk Assessment

- Identify and document the format risks and risk triggers associated with the digital materials, and make *feasible* plans for taking preservation actions, such as storage and format migration.
- Identify "essential characteristics" AKA "significant properties" for different types of files that provide testable success metrics for content fidelity for tools used in format migrations.
- The goal is always the preserve the content of the files. Persevering the full look and feel and user interactions is just not always possible, and that's OK.
- An example is the NARA Digital Preservation Framework, which assesses risk for file formats and identifies the risk mitigation strategies that we use.

Prioritize Basic Levels of Control

- It's deceptively simple to say that an organization has to know what it has, where it is, and who it belongs to when it comes to the preservation of digital objects and files, but that's the place to start.
- The priority should be getting files from wherever they are into a single managed environment if possible - hopefully a single preservation repository. If that's not possible, document the location, level of risk, and who has the responsibility for management and preservation.
- Basic physical controls must be accompanied by metadata, even if it is minimal, such as creator, title, date, provenance, and rights. It's surprising how often we forget about explicitly documenting rights, even if it's to document that there are no restrictions.

Scalable and Flexible Infrastructure

- The Cloud can provide geographical distribution and replication, and is generally easier to scale for processing and storage than on premise data centers. But cloud costs are not simple calculations, as they comprise storage, compute, and egress costs—such as copying files out of the cloud for research requests—when forecasting and budgeting.
- Machine Learning applications can assist with processing and description. Machine learning can be as simple as OCR of scanned text or pattern-matching for PII, or a complex system requiring extensive authority lists and training. But be aware that training machine learning systems is a non-trivial effort.
- Back ups are not archives. Back ups are not preservation. Not all DAMS or Collection Management Systems or Content Management Systems are preservation systems.
- Have a disaster preparedness plan for your infrastructure and systems of record and preservation repository and test those systems for recovery on a regular basis.

Standards and Best Practices

- There are great resources that identify preferred, acceptable, and sustainable file formats, including the Library of Congress Recommended Formats Statement, the Smithsonian Archives Recommended Preservation Formats for Electronic Records, and the NARA Transfer Guidance. The Open Preservation Foundation recently published a draft "International Comparison of Recommended File Formats" as a Google Sheet.
- The Preservation Metadata Implementation Strategy (PREMIS) data dictionary provides guidance for appropriate preservation metadata.
- The Digital Preservation Coalition publishes a Digital Preservation Handbook which provides an excellent overview, as well as their Technology Watch Reports.
- The NDSA (National Digital Stewardship Alliance) has several publications to guide preservation practices.
- The Council on Library and Information Resources (CLIR) has issued dozens of excellent reports, many of which focus on digital collection and digital preservation topics.

Collaboration and Partnerships

- There is a growing community that can provide resources for planning and executing digital preservation programs, share best practices, share access to equipment, and collaborate on shared collection development and preservation projects.
- Community examples include the NDSA, DPC, DCC, OPF, etc. There are services including Hathi Trust, APTrust, Portico, Ithaka, etc.
- There are dozens of mature, open tools for all aspects of preservation workflows, from BagIt for transfers to BitCurator for forensic processing to a variety of systems for processing, description, and preservation: the ArchivesSpace, Museum Space, CollectionSpace, CollectiveAccess, Archivematica, and DuraSpace.

A Note about Web Archiving

- Every single strategy I just discussed applies to web archiving.
- There are guidelines for creating more sustainable/archiveable sites for webmasters, at least in the U.S. Federal space.
- Selection is a form of risk assessment – it's not just how important the content is, but the fact that sites/pages/documents disappear or change without warning.
- Descriptive and structural metadata are vital. It's not enough to capture it and get the files into a managed preservation environment.
- A scalable architecture for crawling, QA, and access is a must with ongoing captures of updates to site and expansions of the use of multiple platforms and multimedia.
- Collaboration is key to ensure the widest possible coverage of web site preservation, whether selecting by geographic area, site type, event, or topic.

It's Not Just Technology, it's People

- Our communities drive what we do, both the creators and the users.
  - They create the digital scholarship that we should preserve
  - They guide us in identifying other digital content to collect
  - They tell us how and where they discover our collections
  - They tell us how they make use of what we collect and preserve

We're Not Failing if We Don't Save Everything

- Don't try to do it all. No single institution can. Do what you can.
- There is no one right way. Do what makes sense for your organization.
- We are succeeding in the larger scheme of things and as a community.

Thank you

Leslie Johnston
leslie.johnston@nara.gov