https://doi.org/10.2352/issn.2168-3204.2020.1.0.1

The Ever-Changing Work That is Digital Preservation

Leslie Johnston; U.S. National Archives and Records Administration (NARA); College Park, Maryland, USA

Abstract

Since the 1960s, digital preservation has transformed from a secondary activity at a select few cultural heritage organizations to a vital international effort with its own best practices, standards, and community. This keynote presentation and paper presents an overview of the changing scope of digital preservation, issues, and strategies for digital preservation in the cultural heritage community.

The Need for Digital Preservation is Not New

Digital Preservation is not new, just like computing in cultural heritage organizations is not new; the introduction of the latter influenced the development of the former.

Most in the cultural heritage community are familiar with the history of the MARC (MAchine-Readable Cataloging) standard for the description of library collections (Seikel & Steele, 2011). Computer scientist Henriette Avram developed MARC in the 1960s in consultation with the Library of Congress; by 1971, MARC was the US national standard, and the international standard by 1973. Large museums started to incorporate mainframe computers into their collection recordkeeping strategies in the 1960s (Johnston, 2012a). The Smithsonian Institution National Museum of Natural History developed SELGEM (Self Generating Master) System in 1965, which it shared with the University of California at Berkeley, the Lowe Museum at the University of Florida, and the Oklahoma Inventory of Ethnological Collections. Fifteen New York-area museums joined forces to explore ways that an electronic index of the Metropolitan Museum's collections could be used beyond the Met, creating a consortium called the Museum Computer Network to create a shared "data-bank" system called GRIPHOS (General Retrieval and Information Processor for Humanities Oriented Studies). Project Gutenberg started marking up texts in 1971 (Manley & Holley, 2012); the Generalized Markup Language (GML) for text was developed in 1970, with the Standard Generalized Markup Language (SGML) standard following in 1983 (Goldfarb, 1999). The distribution of images of collections got its real start in the 1970s: The Museum of Fine Arts, Boston, distributed its first videodisc of 2,000 collection images in 1979. By the mid-1990s cultural heritage were managing and describing their collections in database-driven systems; marking up electronic texts using SGML and HTML (and soon, XML) (Johnston, 2012a, 2012b); publishing and managing access to electronic journals; digitizing their collections; and sharing their expertise and collection on the Web.

That same period of the mid- to late-1990s were also a watershed period for the development of the profession of digital preservation. Robert Kahn and Robert Wilensky published the Handle service specification for the management of digital objects in 1995 (Kahn and Wilensky, 2006). Major research libraries such as Yale, Cornell, and the University of Virginia began large-scale text digitization projects (Johnston, 2012b). The Digital Library Federation was formed in 1995 by twelve academic libraries, the New York Public Library, the Library of Congress, the National Archives and Records Administration (NARA), and the Commission on Preservation and Access (CPA). Brewster Kahle founded the private Internet Archive in 1995, already understanding the potentially short-lived nature of information on the Web.

Don Waters and John Garrett issued a report in 1996 from the Commission of Preservation and Access which called for widespread investment in digital preservation (Waters & Garrett, 1996). The same year, Paul Conway and CLIR issued the report "Preservation in the Digital World," focusing on the need to preserve the files created through digitization (Conway, 1996). In 2000, Stanford University founding its LOCKSS (Lots of Copies Keeps Stuff Safe) program for the distributed preservation of online journals (Dobson, 2003). The same year, the Library of Congress launched its National Digital Information Infrastructure and Preservation Program; the following year the Digital Preservation Coalition was founded in the UK. In 2002, the Consultative Committee for Space Data Systems (CCSDS) published the Recommendation for Space Data System Standards Reference Model for an Open Archive Information System (OAIS) (Consultative Committee for Space Data Systems, 2002). The OAIS model provided a formal reference model for the discussion and development of tools and functions required for digital preservation that is still referenced today. The same year MIT launched its open-source DSpace institutional repository tools (Smith et.al, 2003), followed soon by the release of the Fedora open source digital object management and preservation framework developed by Cornell University and the University of Virginia in 2003 (Payette and Staples, 2002).

Digital Preservation was starting to formalize into a well-understood lifecycle from the creation of objects through processing and ingest to preservation and access, part of an ongoing cycle of review and iteration to ensure that the digital objects continue to be viable and accessible.

Changing Conversations about Digital Preservation

Twenty years ago, a large proportion of our conversations about digital preservation were almost entirely about technology: What standards would we use to build the tools that we needed to store copies of the things that we were worried about? How many copies should we store? Which types of storage media were optimal? What were the appropriate data models for storing those objects? (Arms, 1995; Hedstrom, 1997; Lee et.al, 2002, Levy, 1998) We asked those questions because they seemed discrete and knowable, something we could analyze and answer and test our operations against in an era when we were developing more relevant metrics that defined the necessary attributes and responsibilities of a trusted digital repository (RLG-OCLC Working Group on Digital Archive Attributes, 2002; RLG-NARA Digital Repository Certification Task Force, 2007).

At the same time, those conversations and metrics also revolved around the preservation efforts that supported access, sometimes from a perspective that sometimes engendered a sense of panic as the phrase "Digital Dark Age" appeared frequently in print (Kuny, 1997; Brand, 1999; Wato, 2004; Cox, et.al, 2007). How much would so much storage cost? What were the staffing requirements? Did we really think that we could preserve everything, and when we understood that we definitely could not, what were the selection criteria and collaborations required to ensure that we could preserve as much as possible? What were the digital preservation operational and policy gaps in our organizations given the state of the art in the community? (Hirtle, 2003/2008; Saracevic, 2000) And, when all was said and done, could digital preservation be affordable and sustainable? (Eakin, Friedlander, et.al, 2008; Blue Ribbon Task Force, 2010) The answer wasn't always yes, but the answer could not be no; organizations had to understand the issues and strategies that they could employ because digital preservation was not a required activity, not optional.

Issues that Affect Digital Preservation

What are some of the issues driving digital preservation today? The first is ***Heterogeneity***. No community or organization is going to create, collect, and/or preserve just a single type of born-digital or digitized object. There are literally thousands or tens of thousands of variant versions of file formats over time, and they just keep changing. We cannot identify every legacy format with certainty: Take the .doc file extension. Shorthand for document, it was originally used by WordPerfect as the extension for their proprietary binary text format. In 1983, Microsoft also chose .doc as the extension for their different proprietary binary text format in 1983. Other word processing tools then also allowed users to create files with a .doc extension, which means there are over 30 years of .doc files in existence created by multiple versions of multiple software packages. Or consider the Portable Document Format (PDF). There have been fourteen versions of core PDF—1.0 through 2.0—not to mention subset versions such as PDF/A (Archival). PDF is associated with dozens of Adobe Acrobat releases, stand-alone distiller software, and varying levels of support in hundreds of other applications. Multiply this by every type of business or research function since the 1960s and you will understand the scope of the homogeneity challenge. In another example: the U.S. National Archives first authorized the transfer of born-digital records from federal agencies in 1968, and received its first transfer in 1970; that's 50 years of files just at a single institution. A combination of commercial (current and vintage), open source, and forensic tools are needed to characterize the formats and view and transform the files into sustainable and accessible versions (Kirschenbaum et.al, 2010).

To preserve those files, we first need to be able to read the files off the media that they're stored on. There are dozens of carrier formats—floppy disks, hard drives, CDs, DVDs, thumb drives, tapes, etc, that requires hardware that isn't manufactured or supported by modern personal computer architectures. This also requires a combination of current, vintage, and specialized forensic hardware and driver software (Kirschenbaum et.al, 2010).

Both files and the infrastructures that create them have introduced increasingly ***Complexity***. Born-digital and digitized collections do not exist without context, which must be recorded and maintained; the context includes technical preservation metadata and descriptive provenance metadata. And individual intellectual "items" are increasingly complex, comprised of multi-part or containerized files that require all their components in the correct structure. Consider geospatial (GIS) data files, digital design files, databases, software, and web sites, all of which require all of their parts to accurately render their aggregated content.

The most difficult of all to move into the preservation lifecycle are items or objects or which are created and stored inside systems, never instantiated as files in directories on discrete machines. This is true of business system like personnel systems, case management tools, publishing or document management systems, or web content management systems. This introduces a level or risk where content must be exported from one environment to another, instantiated in new formats, potentially introducing loss of the inherent essential characteristics of the items or their authenticity.

Unsurprisingly, another issue is ***Scale***. As an example, there are thousands of researchers, students, and prominent individuals associated with a single university whose research and personal files will be collected by its archives over time. This is on top of the more traditional library publications, whether physical or digital. Or the over 200 federal agencies creating records. Or the records of every Presidential administration. Or each session of Congress. There is a massive amount of observational data and research datasets created in scientific research that research data preservation policies require that universities and other cultural heritage organizations potentially retain and preserve. The more visible aspect of scale is that there are now huge numbers of files being created by every individual or observational instrument or mass digitization effort every day, and that some types of collections – audio, video, film, email – produce both huge files and huge numbers of files to preserve. They all need to be processed (Green and Meissner, 2005; Johnston, 2015).

The ***Technology*** required for all of these efforts is ever-changing. With heterogeneity comes a wide variety of ever-changing tools and workflows needed to view, process, describe, preserve, and provide access to digital collections. Storage is a major concern when you consider scale and the need for preservation replication; even a "small" collection can take several Terabytes of storage across spinning disk and tape media. With scale also comes stress on local networks and the limiters of moving files using web protocols when operating in the Cloud; most services, web servers, and browsers throttle the amount of bandwidth that can be consumed so no one process can dominate, and set limits on the size of files that can be moved, often to 4 Gigabytes. To work with collections of this size and scale, machines, whether physical or virtual, will require increasingly more storage and memory and higher bandwidth network connections. This will not decrease with time.

The last issue to consider is not technical – it's human: we are serving ***Multiple Communities and Purposes***. There are two key ideas that I always keep in mind: "If it's not accessible, we have not preserved it;" and "We will never be able to guess all the ways that our collections will be used in the future." Both are a reminder that the goal of digital preservation is not just ensuring that we have safe copies of files—of course that's vital—it's that we are preserving our collections for people who need them now, or will need them in the future. Just as there is no single community of creators, neither is there a single, unchanging designated community of users. And new communities will always emerge with new technologies: other machines and web services may soon make up even more of the use of our collections through APIs, but ultimately, it is people guiding and asking for the results of those machine processes. It is well-

known to those trained in collection development theory that we will never know which of our collections will prove the most useful to researchers in the future, or when that day will come, but we must be both collecting the preserving what we can for that future time. We will need to change our own organizations to meet the needs of our collections and our communities.

Strategies for Digital Preservation

What are some of the most successful digital preservation strategies? The digital preservation life cycle starts with the people creating the files, not when the files come over the transom to for us to preserve, so, wherever possible, ***Guidance for Creators*** is extraordinarily valuable. 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 can address file management strategies as well as preferred and acceptable formats and minimum metadata for both the work and for acquisition and preservation.

We must always work to ***Gain Control Over What We Have***. It's deceptively simple to say that an organization has to know what it has, where it is, and who it belongs to—it is not always easy to accomplish but that's the place to start. Inventory and count the files that you have on every box, every server, and every piece of media that make up the collections in all the places and systems where they reside. Match that inventory to whatever metadata you have, no matter how basic, even if it's just the file names, associated custodial unit, file location, provenance, and what you know about the associated rights. These efforts are the necessary basics to work toward a necessary goal of consolidation into fewer storage and systems of record, and, hopefully, into a single preservation environment.

An ***Ongoing Risk Assessment*** is the next step after gaining initial control. Using available tools, characterize the collection file types, even if it's only at the file extension level or MIME type. Use that information to build a collection profile that identifies all the file formats in the collection. If possible, using community resources that identify format sustainability factors, document the risks associated with the format in the collections, and make feasible plans for taking preservation actions, such as storage and format migration, when risk conditions happen (Graf et.al, 2017; Johnston, 2018). The feasible goal for the preservation plans is always to preserve the essential characteristics and content of the files. Persevering the full look and feel and user interactions is just not always possible, and that's OK.

The ability to take preservation actions in accordance with preservation plans requires a ***Scalable and Flexible Infrastructure***. One of the core premises of preservation storage is that multiple copies of files across different storage media and architectures provides the greatest risk mitigation. With increasing numbers, variety, and overall extent of files, local processing resources and on-premise storage will be increasingly difficult to scale up. The Cloud can provide geographical distribution and replication, and is generally easier to scale for processing and storage than on premise data centers (Oliver and Knight, 2015). Whether on-premise, in the Cloud, or a hybrid, one point must be made clear: backups are not archives. Backups are not preservation. Your organization must have managed environment with a disaster preparedness plan for your systems of record and preservation infrastructure, and test those systems for recovery on a regular basis.

Another aspect of a scalable infrastructure is the use of machine learning in the processing of collections. I very carefully do not refer to this as Artificial Intelligence (AI), because even relatively simple machine learning tools can provide a high level of return in processing large collections, especially of textual items. Training a tool to recognize Personally Identifiable Information, named entities (individual and corporate names), and geographic place names can extract valuable descriptive metadata to aid in processing and for access. More complex machine learning tools can be trained to recognize the layout of text on pages to extract fielded metadata. This is not future technology: it is technology that exists right now, and is already in use in multiple cultural heritage organizations (Marciano et.al, 2018).

There is a growing community available for ***Collaboration and Partnerships*** 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. There are dozens of mature, open tools for all aspects of preservation workflows, from transfers to processing and description and preservation, each with communities of practices and support. There are collaborative initiatives for digitization, collection building, virtual collection repatriation, transcription, and authority research (Zarnitz et.al, 2019).

Conclusions

While we still find ourselves discussing digital preservation from a technological perspective—Which formats? Which storage? Which tools? What's the easiest to save given technological constraints? —the work of digital preservation is in some ways more about having the right perspective, or you will find yourself overwhelmed by the issues and challenges and technology. There is no one best technology. There is no perfect workflow. There is no one right way. Do what makes sense for your organization. But you have to do something.

That said, don't try to do it all. No single institution can. Do what you can. We're not failing if we don't save every variety of everything all by ourselves: it must be a community effort.

We are succeeding in the larger scheme of things and as a community. That there is still rhetoric from recent years about a lack of concerted digital preservation efforts and a Digital Dark Age is both puzzling and potentially damaging, as has been pointed out by colleagues (Anderson, 2015). While we can point to high-profile preservation success stories where important resources are saved or recovered, the greatest success is actually that there is a digital preservation community at all, one that is sharing in the development of the community itself, from its standards to its tools and processes, and helping the entire community grow and the profession reach a new overall level of technology use and maturity.

References

[1] D. Anderson. "The digital dark age." Communications of the ACM, vol. 58, no. 12, pp. 20-23, 2015.

[2] W. Arms. "Key concepts in the architecture of the digital library." D-lib Magazine, vol. 1, no. 1, July, 1995: http://dlib.org/dlib/July95/07arms.html

[3] Blue Ribbon Task Force on Sustainable Digital Preservation and Access, and A. Smith Rumsey. Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information: Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Washington D.C.: Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010.

[4] S. Brand. "Escaping the Digital Dark Age." Library Journal, vol. 124, no. 2, pp. 46-48, 1999.

[5] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, Issue 1 (January 2002), http://ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf

[6] P. Conway. "Preservation in the Digital World." Washington, D.C.: Council on Library and Information Science, March 1996: http://www.clir.org/pubs/abstract/pub62.html.

[7] R. Cox. "Machines in the archives: Technology and the coming transformation of archival reference." First Monday, vol 12, no. 11, 5 November 2007: https://firstmonday.org/ojs/index.php/fm/article/download/2029/1894

[8] C. Dobson, C. "From Bright Idea to Beta Test: The Story of LOCKSS." Searcher: The Magazine for Database Professionals, vol 11, no. 2, pp. 50-53, February 2003.

[9] L. Eakin, A. Friedlander, R. Schonfeld, and S. Choudhury. "A selective literature review on digital preservation sustainability." Washington D.C.: Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008.

[10] M. Greene, D. Meissner. "More product, less process: Revamping traditional archival processing." The American Archivist, vol. 68, no. 2, pp. 208-263, 2005.

[11] C. F. Goldfarb. "The roots of SGML: A personal recollection." Technical Communication, vol. 46, no. 1, pp. 75-78, 1999.

[12] R. Graf, Roman, H. M. Ryan, T. Houzanme, and S. Gordea. "A decision support system to facilitate file format selection for digital preservation." Libellarium: journal for the research of writing, books, and cultural heritage institutions, vol. 9, no. 2, 2017.

[13] M. Hedstrom. "Digital preservation: a time bomb for digital libraries." Computers and the Humanities, vol. 31, no. 3, pp. 189-202, 1997/1998.

[14] P. Hirtle. "The History and Current State of Digital Preservation in the United States." Metadata and Digital Collections: A Festschrift in Honor of Tom Turner; Ithaca, NY; CIP (CU Library Initiatives in Publishing); 2008; 121-140.

[15] L. Johnston. "Before You Were Born…Museums had Networks." The Signal, 9 November, 2012: https://blogs.loc.gov/thesignal/2012/11/before-you-were-born-museums-had-networks/

[16] L. Johnston. "Before You Were Born, We Were Digitizing Texts." The Signal, 19 December, 2012: https://blogs.loc.gov/thesignal/2012/12/before-you-were-born-we-were-digitizing-texts/

[17] L. Johnston. "Big Data: New Challenges for Digital Preservation and Digital Services." Jayasuriya, H. Kumar, and Kathryn Ritcheske, eds, Big Data, Big Challenges in Evidence-Based Policy Making. West Academic Publishing, 2015: pp. 27-46.

[18] L. Johnston. "Creating a holdings format profile and format matrix for risk-based digital preservation planning at the national archives and records administration." 15th iPres International Conference on Digital Preservation, Boston, MA, USA: September 24-27, 2018: https://osf.io/ctw3g/

[19] R. Kahn and R. Wilensky. "A framework for distributed digital object services." International Journal on Digital Libraries vol. 6, no. 2, pp. 115-123, 2006.

[20] M. Kirschenbaum, R. Ovenden, and G. Redwine. Digital forensics and born-digital content in cultural heritage collections. Washington D.C.: Council on Library and Information Resources, 2010: http://www.clir.org/pubs/reports/pub149/reports/pub149/pub149.pdf.

[21] T. Kuny. "The digital dark ages? Challenges in the preservation of electronic information." International Preservation News, vol 17, pp. 8-13, 1998.

[22] K. H. Lee, O. Slattery, R. Lu, X. Tang, and V. McCrary. "The state of the art and practice in digital preservation." Journal of research of the National institute of standards and technology, vol. 107, no. 1, pp. 93-106, 2002.

[23] D. M. Levy. "Heroic measures: reflections on the possibility and purpose of digital preservation." Digital Libraries '98: Proceedings of the third ACM conference on Digital Libraries, Pittsburgh, Pennsylvania USA: May 1998, pp. 152-161.

[24] L. Manley and R. Holley. "History of the ebook: The changing face of books." Technical Services Quarterly, vol. 29, no. 4, pp. 292-311, 2012.

[25] R. Marciano, V. Lemieux, M. Hedges, M. Esteva, W. Underwood, M. Kurtz, and M. Conrad. "Archival records and training in the age of big data." Re-Envisioning the MLS: Perspectives on the future of library and information science education vol. 44, pp 179-199, 2018.

[26] G. Oliver, S. Knight. "Storage is a strategic issue: digital preservation in the cloud." D-Lib Magazine vol. 21, no. ¾, 2015.

[27] M. Seikel and T. Steele. "How MARC Has Changed: The History of the Format and Its Forthcoming Relationship to RDA," Technical Services Quarterly, vol. 28, no. 3, pp. 322-334, 2011.

[28] M. Smith, M. Barton, M. Bass, M. Branschofsky, G. McClellan, D. Stuve, R. Tansley, J. Walker. "DSpace: An Open Source Dynamic Digital Repository." D-Lib Magazine, vol. 9, no. 1, January 2003: http://www.dlib.org/dlib/january03/smith/01smith.html

[29] S. Payette and T. Staples T. " The Mellon Fedora Project Digital Library Architecture Meets XML and Web Services." In: Agosti M., Thanos C. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2002. Lecture Notes in Computer Science, vol 2458. Springer, Berlin, Heidelberg.

[30] RLG-NARA Digital Repository Certification Task Force. Trustworthy repositories audit & certification: Criteria and checklist.

Mountain View, CA: Research Libraries Group (RLG), 2007: http://www.crl.edu/PDF/trac.pdf

[31] RLG-OCLC Working Group on Digital Archive Attributes. Trusted digital repositories: Attributes and responsibilities. Mountain View, CA: Research Libraries Group (RLG), 2002: http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

[32] T. Saracevic. "Digital library evaluation: Toward an evolution of concepts." Library Trends, vol. 49, no. 2, pp. 350-369, 2000.

[33] D. Waters and J. Garrett. “Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.” Washington D.C.: The Commission on Preservation and Access,1996: http://www.clir.org/pubs/abstract/pub63.html.

[34] R. Wato. "Challenges of Archiving Electronic Records: The Imminent Danger of a ‘Digital Dark Age’." ESARBICA Journal, vol. 23, pp. 82-92, 2004.

[35] M. Zarnitz, T. Bähr, U. Arning. "Ten Years of Strategic Collaboration of Libraries in Digital Preservation." LIBER Quarterly vol. 29, no. 1, 2019: https://www.liberquarterly.eu/articles/10.18352/lq.10278/

Author Biography

Leslie Johnston received her B.A. in Anthropology and her M.A. in Archaeology, both from the University of California, Los Angeles. Since then she has worked in the cultural heritage community for libraries, archives, and museums, focusing on digitization, born-digital and digitized collection management and preservation, descriptive and preservation standards, and the design and development of digital collection preservation and access systems.