Information Services & Use 40 (2020) 193–199
DOI 10.3233/ISU-200090
IOS Press

Challenges in preservation and archiving digital materials

Leslie Johnston*
Director of Digital Preservation, National Archives and Records Administration (NARA), College Park, MD, USA

Abstract. At its most basic, digital preservation comprises a series of risks, and strategies to mitigate them. And no matter the scale or type of collections, whether born-digital, digitized, or both, the same challenges and risks apply and similar strategies can be employed. This paper identifies a series of common challenges and potential strategies that can be put in place no matter the type or size of collection or collecting organization.

Keywords: Digital preservation, file format, risk analysis, risk mitigation, strategic planning

Digital Preservation is not new. While we might not have called it digital preservation at the time, the National Archives started accessioning and preserving born-digital records fifty years ago, in 1970. Over the past twenty-five years our community conversations about digital preservation became more common, but were initially focused almost entirely on technology: Which types of storage media were optimal? How many copies should we store? What were the appropriate data and metadata models for storing those objects? How would we build and maintain the tools necessary to store copies of our collections [1–4,6,7]? We asked those questions because they seemed discrete and knowable, something we could analyze and answer and test against when we were developing hopefully relevant metrics for the attributes and responsibilities of a trusted digital repository and preservation program [8,9].

As the profession and the programs at our institutions became more mature, conversations shifted. What was necessary not only technologically, but also programmatically to ensure ongoing access to the collections we preserve? Some of the discussions, especially in the public sphere, were couched in a manner that engendered a sense of panic as the phrase "Digital Dark Age" appeared frequently in print [10–13]. That panic, whether right or wrong, drew attention to the issues and resulted in additional public funding for a hoped-for collaborative, national infrastructure. When planning for infrastructure to preserve "everything," how much would so much storage cost, even if it was distributed? What were the staffing requirements? Did we really think that we could preserve everything, and when we collectively 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 [14,15]? And, when all was said and done, could digital preservation be affordable and sustainable

[*] Tel.: +1 301 837 3625; E-mail: leslie.johnston@nara.gov.

[16,17]? Of course the answer wasn't always "yes", but the answer could never be "no"; organizations had to instead develop a spectrum of strategies and infrastructure that they could employ.

1. 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, perhaps tens of thousands, of variant versions of file formats going back to the mid-twentieth century, and the growth of formats will not stop. We cannot identify every legacy format with certainty: consider 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. Other word processing programs also allowed users to select their own extensions, which meant that there are over thirty years of .doc files in existence created by multiple versions of software packages from multiple operating environments, not to mention thousands of other random extensions for word processing files. Or consider PDF (Portable Document Format). There have been over one dozen versions of core PDF - 1.0 through 2.0 - not to mention numerous specialized subset versions such as PDF/E (Engineering) or PDF/A (Archival). PDF is associated with dozens of Adobe Acrobat releases dating back to the mid-1990s, stand-alone distiller software bundled into other applications, and varying levels of support for PDF creation and viewing in hundreds of other applications.

Multiply this by every possible type of research and business activity for fifty to sixty years and you will understand the scope of the challenge. 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 fifty years at a single institution, comprising a collection of more than two billion born-digital files and growing. A combination of commercial (current and vintage), open source, and forensic tools are needed to characterize the formats where possible, to view the files to confirm their content and describe them, and to transform the files into sustainable preservation formats and accessible public use versions [18].

In order to preserve those files, the first challenge is the ability to read the files from the media on which they are stored. How many of us still have files on 3.5" floppies or ZIP cartridges that require hardware that we no longer have? There are dozens of carrier formats - floppy disks, hard drives, CDs, DVDs, thumb drives, tapes, etc. - that require hardware that isn't manufactured or supported by modern personal computer architectures and software that can no longer be found online, even if the original manufacturer still exists. This also requires a combination of current, vintage, and specialized forensic hardware and driver software [18].

Both files and the infrastructures that create them introduce the second issue - increasing *Complexity*. Born-digital and digitized collections do not exist without context, which must be recorded and maintained. The context includes the provenance of the files, their original arrangement and intellectual relationships, preservation metadata, and descriptive metadata. And individual intellectual "items" are increasingly complex objects, comprised of multi-part compound or containerized files that require all their components to be retained in the correct, documented 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. In some cases, one cannot convert file formats or change the structure because content integrity and functionality will be altered - even just migrating the formats may lose vital functionality with which to interact with the content.

The most difficult of all to move into the preservation lifecycle are items or objects which are created and stored inside systems. They may never be instantiated as discrete files in directories on machines, or may exist in a hybrid state where descriptive, technical, and structural metadata is in a system and some linked component files are elsewhere. This is true of business systems such as personnel systems, case management tools, publishing or document management systems, or web content management systems. This introduces a level of risk where content must be exported from one environment to another and instantiated in one or more 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 any given university whose research and personal files will be collected by its archives alongside its corporate records. This is on top of the more traditional library monograph and serial publications, whether physical or digital. As another exaple, the National Archives record collecting scope includes the more than four hundred departments, agencies, and sub-agencies in the federal government that are creating records, and the records of each Presidential administration and session of Congress. There is a massive amount of observational data and countless research datasets created as direct research products or data collection efforts that universities and other cultural heritage organizations must potentially retain and preserve due to federal and local research data preservation policies. Another visible aspect of scale is that there are now huge numbers of files being created by every individual or observational instrument or 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 require processing, sometimes just to appraise whether they will be preserved at all [19,20].

The third issue is that the *Technology* that is required for all of these efforts is changing nonstop. 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 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 a small number of Gigabytes. Larger file types are often broken into several smaller files to be moved, or are compressed in ways that can be risky to data integrity and authenticity when the files are unpacked. To work with digital collections of any size and scale, machines, whether physical or virtual, will require increasingly more storage and memory, faster and more processors, more storage, and higher bandwidth network connections. This will not decrease over time.

The last issue to consider is not technical - it is human. We are serving *Multiple Communities and Goals*. There are two key concepts 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 in which 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 is 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 uniform and unchanging designated community of users. And new communities will always emerge with new technologies; for example, other machines and web services may soon make up even more of the use of our collections through APIs and Linked-Data than human researchers. But ultimately, it is people guiding and asking for the results of those machine processes. It is well-known to those who trained in collection development theory that we will never know which of our collections will prove to be the most useful to researchers in the future, or when that day will come, so we must be both collecting broadly and 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.

2. 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 to our organizations to preserve them, 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 for the full range of business, personal, creative, or research endeavors. 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 - this mandate is not always easy to accomplish. 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, provenance, and what you know about the associated rights. That last is surprisingly important, as it affects how and if you can provide access and the ability to reuse/remix the items. 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. That's the ultimate goal for reducing risk.

An *Ongoing Risk Assessment* is the next step after gaining initial control. Using available tools, characterize the collection file formats, 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. You may not be able to characterize or validate the formats at a highly-granular level, such as Microsoft Works Database for Window 3.0a, but even knowing if you have databases versus text files versus images is helpful in risk assessment. If possible, use community resources that identify format sustainability factors to document the risks associated with the format in the collections, and create feasible (not aspirational) digital preservation action plans for taking preservation actions, such as storage and format migration, when risk conditions happen [21,22]. Your process may be manual and not automated, but that's fine, as long as you're doing it. The hoped-for goal for the preservation plans is to always preserve the essential characteristics and content of the files to the fullest extent possible in every storage or format migration. Persevering the full look and feel and user interactions is just not always possible, and that's also 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, combined with geographic distribution, provides the greatest risk mitigation. In the past we would have said one copy on spinning disk, one on a different type of media in the same or a different location, and a backup copy somewhere else, often on tape. With increasing numbers, variety, and the overall extent of files, local processing resources and on-premise storage will be increasingly difficult to scale up to cover multiple copies. The Cloud can provide geographical distribution and replication, and is generally easier to scale for processing and storage than on-premise data centers [23]. This is not to say that there are no inherent risks in the Cloud; you are trusting preservation services to a corporate or non-profit entity that could disappear at any time, and whose storage over which you may have little to no direct control. At scale, though, the tiered storage costs and replication are an attractive option. Whether on-premise, in the Cloud, or a hybrid, one point must always be made clear: backups are not archives. Backups are not preservation. Your organization must have a managed environment with a disaster preparedness plan for your systems of record and preservation infrastructure, and must test those systems and media 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 collections 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 from file headers or record content - OCR'ed text from page images, full-text documents, datasets, audio or video tracks, or even still images -will aid in processing and in providing 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 [24].

Not all strategies are technical. 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 with communities of support 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, authority research, and digital preservation storage. No institution should assume that is it on its own or that it shoulders all responsibility for digital preservation [25,26].

3. Conclusions

While we still find ourselves pondering digital preservation from a technological perspective - Which formats? Which storage? Which tools? What is the easiest to save given technological constraints? - the work of digital preservation is in some ways more about having the right perspective, otherwise 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. However it is both puzzling and potentially damaging that there is still rhetoric from recent years about a lack of concerted digital preservation efforts and a Digital Dark Age [26]. It is equally puzzling that public awareness of the need for personal digital preservation efforts is so low despite the fact that the average person generates Gigabytes and Gigabytes of content every year [27]. And it can be enraging that there are frequently articles in the general media and technology press expressing surprise that libraries, archives, and museums have a role in these preservation activities or have been doing this for decades.

While we can point to high-profile preservation success stories where important resources are saved or recovered, the greatest success is actually that a digital preservation community exists - 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] W.Y. Arms, Key concepts in the architecture of the digital library, *D-lib Magazine* **1**(1) (1995), http://dlib.org/dlib/July95/07arms.html, last accessed on June 6, 2020.

[2] P. Conway, *Preservation in the Digital World*. Council on Library and Information Resources, Washington DC, 1996, http://www.clir.org/pubs/abstract/pub62.html, last accessed on June 6, 2020.

[3] D. Waters and J. Garret, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Council on Library and Information Resources, Washington DC, 1996, http://www.clir.org/pubs/abstract/pub63.html, last accessed on June 6, 2020.

[4] M. Hedstrom, Digital preservation: A time bomb for digital libraries, *Computers and the Humanities* **31**(3) (1997), 189–202.

[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, 2002, http://ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf, last accessed on June 6, 2020.

[6] 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* **107**(1) (2002), 93–106.

[7] D.M. Levy, Heroic measures: Reflections on the possibility and purpose of digital preservation. in: *Proceedings of the third ACM Conference on Digital Libraries,* 1998, pp. 152–161.

[8] RLG-NARA Digital Repository Certification Task Force. *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Research Libraries Group (RLG), Mountain View, CA, 2007, Available at: http://www.crl.edu/PDF/trac.pdf.

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

[10] T. Kuny, The digital dark ages? Challenges in the preservation of electronic information, *International Preservation News* **17**: (1998), 8–13.

[11] S. Brand, Escaping the Digital Dark Age, *Library Journal* **124**(2) (1999), 46–48.

[12] R. Wato, Challenges of archiving electronic records: The imminent danger of a “Digital Dark Age”, *ESARBICA Journal* **23**: (2004), 82–92.

[13] R.J. Cox, Machines in the archives: Technology and the coming transformation of archival reference, *First Monday* **12**(11) (2007), Available at: https://firstmonday.org/ojs/index.php/fm/article/download/2029/1894.

[14] P.B. Hirtle, The history and current state of digital preservation in the United States. in: *Metadata and Digital Collections: A Festschrift in Honor of Tom Turner*. CIP (CU Library Initiatives in Publishing), Ithaca, NY, 2008, pp. 121–140.

[15] T. Saracevic, Digital library evaluation: Toward an evolution of concepts, *Library Trends* **49**(2) (2000), 350–369.

[16] 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*. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Washington DC, 2010.

[17] *A Selective Literature Review on Digital Preservation Sustainability*. Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Washington DC, 2008.

[18] M. Kirschenbaum, R. Ovenden, G. Redwine and R. Donahue, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Council on Library and Information Resources, Washington DC, 2010, http://www.clir.org/pubs/reports/pub149/reports/pub149/pub149.pdf.

[19] M. Greene and D. Meissner, More product, less process: Revamping traditional archival processing, *The American Archivist* **68**(2) (2005), 208–63.

[20] L. Johnston, Big data: New challenges for digital preservation and digital services. in: *Big Data, Big Challenges in Evidence-Based Policy Making,* West Academic Publishing, 2015, pp. 27–46.

[21] R. Graf, 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* **9**(2) (2017), 267–274.

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

[23] G. Oliver and S. Knight, Storage is a strategic issue: digital preservation in the cloud, *D-Lib Magazine* **21**(3/4) (2015), Available at: http://mirror.dlib.org/dlib/march15/oliver/03oliver.html.

[24] 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, 44, 2018, pp. 179–199.

[25] T.O. Walters and K. Skinner, Economics, sustainability, and the cooperative model in digital preservation, *Library Hi Tech* **28**(2) (2010), 259–272.

[26] M. Zarnitz, T. Bähr and U. Arning, Ten years of strategic collaboration of libraries in digital preservation, *LIBER Quarterly* **29**(1) (2019), Available at: https://www.liberquarterly.eu/articles/10.18352/lq.10278/.

[27] D. Anderson, The digital dark age, *Communications of the ACM* **58**(12) (2015), 20–23.

[28] M. Condron, Identifying individual and institutional motivations in personal digital archiving, *Preservation, Digital Technology & Culture* **48**(1) (2019), 28–37.