Small Steps Down the Superhighway:

Preserving Born-Digital Artifacts and Creating Digital Archives

Dana Rigg

Old Dominion University

LIBS 610: Preservation Management in Libraries and Archives

Professor Jessica Ritchie

December 16, 2022

Small Steps Down the Superhighway: Preserving Born-Digital Artifacts and Creating Digital Archives

Digital preservation cannot wait. Archival institutions may need to preserve digital objects either as backups to their physical artifacts or as archival artifacts in their own right, as outlined by the institution's collections policy. Institutions may wish to preserve digital artifacts as a public service or to increase access. Individuals may need to preserve their own digital artifacts for legal or personal reasons, and may look to preservation institutions for guidance. All of these are good reasons to begin a digital archive. By taking small steps from understanding digital media and its inherent vice and preservation challenges, to thinking through ethical implications, to keeping in mind key tips for preliminary planning, we can start down the road of starting a digital collection in earnest.

The NC State University Libraries have identified three major types of digital artifact (De Klerk, 2018). Understanding these three types offers a helpful place to start, as they each involve different but overlapping preservation challenges.

First, there are digitized items. These are digital copies, images, or simulated representations of physical artifacts. These may be a scan of a handwritten letter, a photograph of a piece of artwork, or a 3D rendering of an antique bureau. In all cases, a digitized item is something that was originally created in a physical form and was later copied digitally.

The next level of digital artifact is the born-digital artifact. As OCLC defines them, ". . . born-digital resources are items created and managed in digital form" (Erway, 2010, p. 1). This is a document, photograph, data set, or artwork that began as a digital file and not a physical object. Consider a Microsoft Word document on a hard drive, an original computer game stored on a floppy disk, or a digital photo on an SD card. Although these may have physical containers, the

information artifact itself only exists digitally and must be accessed and displayed digitally. This is one kind of born-digital artifact.

Another type of born-digital artifact is the web-based artifact. This third kind of digital artifact is distinct from other born-digital content because a web-based artifact is hosted on the internet, and so is not attached to a physical container at all. Something like a Google Doc, an email thread, a university website or a personal blog would be considered a web-based artifact.

Any of these three types of digital content – digitized, born-digital and web-based – may fall under an institution's existing collection development policy. They may represent a new challenge for archivists who may not have the expertise or the infrastructure ready to tackle preservation of a new format. Nonetheless, failure to act quickly may lead to irreversible loss.

This loss can take several forms. Born-digital files tied to physical components may have the most obvious areas of loss. One of these is hardware obsolescence. This occurs when "due to technological advancements, very few, if any, of the computers or peripherals needed to access the files still exist, and they are difficult to repair or replace" (Erway, 2010, p. 3). Consider similar challenges with A/V technology like microfilm readers or 8-track tape players – the same access difficulty can happen to floppy disk drives or other computer-based access devices.

Born-digital documents can also experience software obsolescence. This occurs when the software or operating system needed to read a file is no longer in service (Erway, 2010). If a born-digital artifact is saved in an obsolete file type, it is functionally worthless.

Similarly, the media itself can become unusable, through corruption over time or through damage to the media container itself. If you've experienced a CD skipping, you've experienced media failure. The Digital Preservation Coalition (2016) notes that "storage media is commodity [sic] product and tends to have a reasonably short lifespan. Most hard disks tend to have a

reliable lifetime of around 5 years" (p. 3). Additionally, any information item stored physically is in danger from disaster such as fire or flooding. Even if a file is born-digital, it can still be permanently destroyed if not backed up.

Digitized and web-based files can also be lost or damaged. Digital media has its own inherent vice, bit rot, that can affect digital files over time. As defined by the DPC, "bit rot refers to the loss of data due to the small electronic charge of a bit, or alternatively, by cosmic rays or other high energy particles" (p. 3). Risks of media degradation like this can be total inaccessibility of the file, or it can be more subtle, such as a warping, pixelating or fuzzing of the appearance of an image. The DPC (2016) reports of one case,

> "some of the newspaper pages that were . . . damaged looked fine until you zoomed in, and [then] they became fuzzy. Although the bitstream was damaged, the viewer software did it's [sic] best to render the image without informing the user. Things are not always as they seem!" (p. 4)

This type of loss is insidious in an archival setting, because once this kind of bit rot has set in unnoticed, it is impossible to repair and any copies made after that point will be equally damaged.

A type of loss that is similarly subtle and irreparable is the loss of authenticity. As NC State University notes, "another important ethical part of our digital program is ensuring that the digital records we create and preserve are authentic. Born digital materials in particular are easily modified or corrupted" (De Klerk, 2018, para. 11). Tampering or alteration to a digital record may not leave a trace. Without archival protection, digital files can become worthless as historical artifacts through the loss of authenticity.

Finally, without preservation, web-based artifacts can vanish. The Brooklyn Public Library reports that the average lifespan of a website is only 100 days (Bowers-Smith, 2021). Whether from deliberate deletion, domain failure, accident, or other issue, a vanished web-based artifact is knowledge lost. Even when websites remain, they frequently suffer broken links which direct to moved or deleted pages. This can make it difficult to preserve a cohesive historical record. This short-lived and fast-moving area demands quick action and new strategies from archivists.

Archivists and other preservation professionals are aware of the need for action. Light (2010) entreats, “don't wait until it's too late to take action to preserve your born digital materials. The future of this kind of scholarship depends on your good management of your files. . . ” (p. 13). Post (2021) is optimistic about the ability and benefits of archivists taking an active, innovative part in digital preservation, saying, “for digital cultural materials specifically, post-custodial approaches provide frameworks and methods for proactively stewarding fragile digital objects that are at risk of being lost to obsolescence long before they pass over the custodial threshold” (p. 11). In short, this work cannot wait, and it cannot be left to others.

Institutions of all sizes and specialties are approaching this task with an innovative attitude. The National Library of Australia launched its national digital collection in 2019 by expanding the legal deposit requirement, which mandates that a copy of all published works in Australia be sent to the National Library. The mandate now extends to all published ebooks, creating a complete collection of Australian-published digital books (Torney, 2019). The Brooklyn Public Library launched its web-preservation program in 2017, where it uses tools from the Internet Archive to preserve snapshots of local Brooklyn news and culture websites (Bowers-Smith, 2021). The curation team of Richard Rorty’s archive at the University of

California, Irvine worked to migrate and preserve his born-digital documents along with his papers, creating a hybrid archive which fully represents his work. The curators of Salman Rushdie's archive at Emory University went one step farther, choosing to emulate his original Macintosh computer environment to most accurately and immersively preserve his born-digital work (Light, 2010). Universities of all sizes are beginning digital collections to preserve the work of students and staff along with born-digital official documents that their collection policy demands they preserve (Herzinger et al., 2021; Griffin, 2015). Even at the individual level, archivists are working to inform people of steps they can take to organize and preserve their own born-digital documents (Light, 2010). Possibilities for digital preservation exist at all levels, even if it may seem intimidating.

So after establishing the necessity of digital preservation on an archival level, where do we go from here? As with any other archival collection, there are things to consider at the beginning of the process. Some important ethical areas include privacy, copyright, access, and curation.

Privacy concerns for archives of born-digital materials are just as essential as traditional archival materials, if not more so. NC State University Libraries advise that "creators' expectations of privacy, as well as the people who may be mentioned in the materials, are something worth thinking about when deciding whether or not to provide online access" (de Klerk, 2018, para. 8). This applies to born-digital artifacts like social media posts, whose creators intended it to be received publicly, but also to digitized records, whose creators may not have. There may exist a happy middle ground by limiting access to some materials while reserving others; the Rorty archivists found this by setting up a virtual reading room requiring users to consent to conditions of use, and publishing metadata about the collection

to aid discovery by outside researchers (Light, 2010, p. 9). Solutions to privacy concerns in the digital space must be sought, and may require creative solutions.

Copyright concerns are similar to privacy concerns in that they can become even more hard to define when dealing with born-digital documents. Born-digital donations, such as the Rorty archive received along with the philosopher's papers, may have similar copyright considerations to traditional media. The rights of publication may be similar when digitizing or publishing the born-digital files along with the physical, but it is necessary to be clear about the rights of the institution from the outset (Light, 2010, p. 8). Some other born-digital cases may be less clear-cut. Like traditional media, it can sometimes be difficult to determine the copyright holders for a piece of digital media. The NC State University Libraries advise only putting materials online in a low- or no-risk situation, deciding on a case-by-case basis and offering a route for recourse if the rights-holders do want to request removal (de Klerk, 2018, para. 8, 10). Responsible stewardship requires archives to protect the rights of digital rights-holders as well as the institution.

Access is an area that, with digital media especially, combines the concerns of privacy and copyright with the needs of researchers. Rorty's archivists encountered this when they chose to restrict access to a digital reading room in order to preserve privacy and respect copyright claims (Light, 2010). The NC State University Library staff present their situation as a series of trade-offs, requiring them to choose between thoroughness of description and the greater quantity of digital materials that can be preserved with less description (de Klerk, 2018, para. 19). Their priority is the maximum possible access for all individuals, with a focus on reaching marginalized communities. Each institution has chosen a level of access to their digital collection that accounts for the privacy concerns and the mission scope of their collection.

Possibly the most important area to consider before beginning a digital archive is curation. The quantity of born-digital and web-based materials is vast, and archival resources and staff are limited. As Light (2010) puts it,

> The reality is that most archival repositories . . . don't have the staff, space or resources to become a technological museum with many kinds of equipment, operating systems, and software packages. As an archivist, I faced long ago that I don't have the resources to save everything, and hard choices have to be made daily about what to carry forward with us from the past. (p. 5)

Archives may find themselves in a race against time between vanishing web-based artifacts and deteriorating born-digital files. Some archives may choose to focus on one or the other, and some may be forced to prioritize. Based on an assumption that the born-digital files on physical media were already backed up, archivists at the University of Louisville, KY, decided to "start in the present moment and keep pace with . . . incoming born-digital collections. As we had time, we would shift focus to things like rogue media—those CDs and floppy disks hidden away in waiting-to-be-processed collections" (Herzinger et al., 2021, p. 360). To formulate an effective plan, it is important to define the scope of the digital collection, determine the priority for preservation, and proceed from that understanding.

Once we understand the risks inherent in our digital media, the urgency of the need to preserve born-digital documents for the benefit of the historical record, and the ethical considerations that affect digital archives, we can begin to move toward practical preservation. Regardless of the scope or nature of your digital collection, there are some general tips that will provide a helpful starting point.

Tip 1: Survey the literature. The information you need will vary greatly depending on the types of digital documents you need to preserve. Learning from others will save you time and stress. Hertzinger, et al. at the University of Louisville (2021) discuss at length the lessons they took from the sources they surveyed as they were starting their digital collection, including defining the initial scope and pitfalls to avoid. Of their research, they said, "these case studies provided examples of specific tasks that could inform aspects of the digital preservation program we wanted to build" (p. 354). OCLC also recommends that you "find others who are working on similar challenges. There may be already existing standards, tools, or procedures used by another community, such as law enforcement or gamers" (Erway, 2010, p. 4). Researching the specifics of your unique collection will put you on the right track.

Tip 2: Begin with a digital preservation policy. Wrestling with ethical issues and learning from the records of similar collections should lead to a firmer footing for your digital collection. A policy will further define the scope of the collection, help settle priority, and ensure that the demands of the collection will not get out of hand.

Tip 3: Use the right tools. Products exist to help archive web pages (Bowers-Smith, 2019; de Klerk, 2018), to preserve integrity by using checksums (DCP, 2016), to scan for and censor sensitive data to protect privacy (de Klerk, 2018), to make documents full-text searchable (de Klerk, 2018) and to host the catalog (Herzinger et al., 2021), as a very basic start. Don't assume you need to build anything from the ground up.

Tip 4: Be realistic. You don't need to have a perfect collection right away. As long as your records are preserved and stable, you have time to work out issues of budgetary constraints, access, and presentation. Herzinger et al. (2021) point out how

> [all] too often, Kenney and McGovern argued, curators focus on systems to solve the digital preservation challenge, prompting many of us to believe that digital preservation is either completely developed or not at all. This leads curators to reject the reality that an early stage program at one institution may be a fully developed program at another institution." (p. 358)

They emphasize that as long as digital artifacts are well reproduced and are stored securely, that may meet your needs (p. 357). Be realistic about your needs and expectations.

Tip 5: Don't hesitate. Digital preservation is still in its early years, so the lack of established norms may give some archivists pause. But the fast-moving, constantly changing, and highly unstable nature of born-digital artifacts means that without swift action, we will lose valuable historical records. As Light (2010) described their situation, **"**we had to act now or risk forever losing [digital] contents. We didn't wait for clear best practices or standardized methods from the profession. We simply did our best to preserve and provide access to the materials with our available resources" (p. 4). Others echo this call to action: Herzinger et al. (2021) "encourage curators to push through their paralysis and to either launch or incrementally grow and improve their digital preservation programs" (p. 368). Diana Bowers-Smith (2021) of the Brooklyn Public Library says, "my advice would be to just go for it. The best way to learn is by diving in" (para. 12). By starting small and taking practical steps, a valuable digital collection can grow.

Born-digital artifacts need to be preserved, and we must not wait to start. Their ephemeral nature and inherent vice mean that immediate action is needed if we want to avoid gaps in the historical record. By taking small steps, archivists, curators, and information professionals can lead the way and create innovative solutions to preserve our future.

References

Bowers-Smith, D. (2021, February 23). *Web archiving at BPL: Saving Brooklyn's web content one URL at a time.* Brooklyn Public Library. https://www.bklynlibrary.org/blog/2021/02/23/web-archiving-bpl-saving

De Klerk, T. & Serrao, J. (2018, June 1). *Ethics in archives: Decisions in digital archiving.* NC State University Libraries. https://www.lib.ncsu.edu/news/special-collections/ethics-in-archives%3A-decisions-in-digital-archiving

Digital Preservation Coalition. (2016, June 15). *Just keep the bits: an introduction to bit level preservation.* https://www.dpconline.org/docs/miscellaneous/events/2016-events/1662-bit-preservation-gettingstarted/file

Erway, R. (2010, November). *Defining "born digital."* OCLC Research. https://www.oclc.org/content/dam/research/activities/hiddencollections/borndigital.pdf

Griffin, K. (2015, Fall). *Web archiving collection policy.* The University of Wisconsin–Madison Archives. https://cms.library.wisc.edu/archives/wp-content/uploads/sites/21/2016/01/Web-Archiving-Policy.pdf

Herzinger, K., Daniels, C., & Fox, H. (2021). Preservation not paralysis: Reflections on launching a born-digital preservation program. *Collections: A Journal for Museum and Archives Professionals, 17*(4), 347-371. https://doi.org/10.1177/1550190620978221

Light, M. (2010, May 14). *Designing a born-digital archive.* University of California, Irvine. https://escholarship.org/uc/item/8wf5w4nk

Post, C. (2021). The Art of Digital Curation: Co-operative Stewardship of Net-Based Art. *Archivaria*, *92*, 6-47. https://archivaria.ca/index.php/archivaria/article/view/13813

Torney, K. (2019, August 16). *Australian libraries join forces to build national digital collection."* Australian Book Review. https://www.australianbookreview.com.au/features/book-talk/464-book-talk/5744-australian-libraries-join-forces-to-build-national-digital-collection-by-kate-torney