ISSN 2324-7878 (online)
palrap.org

PaLA | College & Research Division (CRD)

Pennsylvania Libraries: Research & Practice

Commentary

Format Agnostic Archival Processing

Using One Standard for All

Kari May

Kari May is the Digital Archives & Preservation Librarian, Archives & Special Collections Department, University of Pittsburgh, karimay@pitt.edu

Pooling guidance and examples from nationally and internationally known groups including the Digital Preservation Coalition, University of California Libraries, and OCLC, this article supports processing archival assets with a format agnostic perspective to promote a unified standard for all materials. Format agnostic processing can embrace and align digital and physical assets under one expectation for basic outputs regardless of their formats. In mixed material collections, digital assets are frequently seen as requiring a disassociated set of tools and knowledge and approached as unconnected to the physical assets. The format agnostic perspective can eliminate the tendency toward varying levels of processing and differing descriptions standards complicating the discovery of collection materials due to the varying formats of its assets. It can also support timely processing and preservation of digital assets to decrease the risk of data loss. Format agnostic processing would enhance discoverability while minimizing the loss of digital information.

For nearly 20 years, there has been an unfortunate narrative surrounding the archival processing of digital archival assets that has segregated it into a world of its own. It takes special knowledge and tools. It takes experts that know how to handle different types of records in different formats. It requires specialized handling and storage. Sure, but these same requirements are encountered when working with physical assets, right? The specifics may be different-it takes Adobe Photoshop to trim a photo instead of an X-ACTO blade and an archival information package for storage instead of a Hollinger box. Yet, paper-based preservation techniques such as photocopying and microfilming are directly comparable to the digital preservation techniques of migration and transformation ("Preservation Methods and Techniques," n.d.). Essentially, there is not anything new here. For this very reason, those in organizations with existing skills in either information management or information technology are well placed to build on and apply these skills to digital preservation activities ("Preservation Issues," 2015). The focus for processing digital content should not be the fact that it is digital but that it is a record of historical value. During the pandemic, many saw the once obvious separation of physical and digital collections begin to fade. As archives move forward, perhaps stepping toward

DOI 10.5195/palrap.2022.269

archival processing that embraces all assets regardless of their formats with the same expectation for basic outputs - a format agnostic perspective- can promote a more unified standard that would enhance discoverability and access.

A format agnostic perspective in processing would encourage the more thorough response to research inquiries since the basic processing of all assets would be equal and standardized ensuring access to these assets no matter what form they take. Consider this scenario: a researcher requests access to an email account in your collection because messages are known to discuss a specific project. Sure, the format agnostic perspective would not affect the quick response (if that collection is open) of just producing the collection as requested. However, the more thorough response would be to ask the date range of interest and include all correspondence available be it email, text, formal memo, or handwritten letter. This way existing physical items could be offered along with the digital to provide the best opportunity for the topic to be fully understood and assessed.

From a high-level perspective, format agnostic processing is already initiated in institutions that preserve archival objects such as photographs, audio-visual media, or artworks alongside books, manuscripts, and newspapers. The varied contents of mixed-media collections, despite the obvious difference in preservation needs, have long been accepted and processed by archives. These collections have fallen within an institution's collection and archival appraisal policies which have determined that the objects have historical value. The introduction to Kentucky State Archives' Electronic Records recognizes this when it says, "Records management standards and principles apply to all forms of recorded information, from creation to final disposition, regardless of the medium in which the records are created or stored. Information in electronic recordkeeping systems must remain accessible to the appropriate parties, until all the legal, fiscal, and administrative retention periods have been met, in the same way as physical records" (Kentucky Department for Libraries and Archives, 2002). The items' formats do not stand alone as the reason for their value. Their provenance and content, authenticity and reliability, order and completeness, condition, intrinsic value, and the costs to preserve them all play a role. The same can be said about digital assets. Just like their written-record companions, digital assets can have archival value on their own or as part of a collection ("Appraisal," n.d.).

We must also consider the increasing prevalence of born-digital material that is making siloed approaches for their processing less and less feasible. Siloes can too often "trap" processes within a system or even vocabulary used only by a specific group and thereby diminish the potential for productive communication throughout workflows. For example, a ZIP disk is found in a collection and taken to the digital preservationist. If the processing of digital assets is siloed, the digital preservationist will extract the content, complete any technical steps required, and preserve that content with minimal, if any, input from collection archivists. The result will be a collection containing assets processed from different standards receiving various levels of attention and offering inconsistent access points. In 2021, the University of California Libraries presented their revised guidelines for efficient archival processing in an OCLC webinar entitled Works in Progress Webinar: Holistic approaches to born-digital appraisal and accessioning -- revising the UC Guidelines. They used the term "holistic," but these guidelines clearly state that siloes need not happen (Arroyo-Ramirez et al., 2021). Processing born-digital assets is still archival processing and can be fully operationalized within archival programs (Dundon et al., 2020). A format agnostic perspective would set baseline requirements for all assets and create a universal language for the aspects, activities, and outcomes of archival processing. Just as the work of AV specialists and photo archivists has been woven into the work done for paper-based materials, archivists who are responsible for digital collections must also be able to talk with fellow archivists about institutional best practices for acquiring, processing, and providing access to digital content. Interweaving processing considerations and management strategies for born-digital content into institutional standards will lay the groundwork for these communications (Dundon et al., 2020).

Key to continued access to any stored assets is organizing and describing them to ensure that all staff whether today or in the distant future can find, access, understand, and use them (The Three Essentials of Digital Preservation Part 1: File Storage, 2018). It can be important to remember that selection, appraisal, and disposal are significant components in any digital management activity ("Preservation Issues," 2015). These are tasks that require the sustained attention and professional judgement of archivists to analyze and synthesize the content of one or more archival assets (Weber, Chela Scott, 2020). Although artificial intelligence (AI) and machine learning are improving, they are still imperfect. Scanning text for recurring terms and analyzing where they fall can be telling at times but let us never forget that an entire book could be written about cats that never actually uses the word "cat." The English language is wonderfully flexible, and humans are endlessly creative. AI has come a long way in object recognition in photographs, but, again, interpreting those photos is something else. So, at this time, nothing can replace human involvement in the appraisal of digital archival items and the determination of their historical value. Here format agnostic standards for processing would be most beneficial. Although working with these assets will require using some programs and possibly some automated processes to retrieve and view the content, the organization and description of these assets need not differ from their physical counterparts. Baseline agnostic standards would clearly define the expected outputs for both of these tasks.

With the expected outputs defined, it would be easy to believe that processing digital assets can be completed with the height of efficiency via automated tools. However, the OCLC blog Time Estimation for Processing Born-Digital Collections found that the amount of work it will take to achieve a particular level of processing for digital assets depends on the existing level of organization and understanding of the collection when it arrives at an institution, and these levels of processing do not cleanly translate to standard time estimates. Both human and machine capacities impact the time required to address the needs of a collection, and both may differ significantly from one another (Weber, Chela Scott, 2020). This means the idea that digital content can be set aside "for later" is an unfortunate belief.

Which leads to the question: Why should a new perspective be considered? All this talk about the potential for a format agnostic perspective may seem rational, but the change will most likely be fraught with challenges and many shifts in thoughts and habits ingrained from years of practice. Ironically, the change is needed because the digital assets are much more fragile than many realize.

Perhaps it's because our society encounters so much technology every day. Perhaps it's because digital assets feel so separate, invisible. Whatever the reason, a tendency does exist for the historical value of digital assets to be overlooked. At the same time, society's familiarity with tech leads to a belief that digital assets are far more reliable than they truly are. The most unfortunate outcome of this is that technology moves so quickly toward obsolescence that allowing digital assets to sit and wait all too often drops that information into a digital black hole. Even with the storage media in your hand, what it holds is lost all the same.

In 2001, it was said that 'Preserving digital information is plagued by short media life, obsolete hardware and software, slow read times of old media, and defunct Web sites' (Chen, Su-Shing, 2001). In 2020 John McClellan Marshall's article The Modern Memory Hole: Cyberethics Unchained showed that, almost twenty years later, the same perspective was being stated and the difficulty in accessing information held in antiquated- much less obsolete- hardware and software continues to grow. An example offered is the inability of a laptop manufactured and programmed after 2010 to read any document created in WordPerfect unless that laptop is extensively augmented and sometimes even that will not suffice. The amount and type of information "lost" in situations like this is simply unknowable, but the steps toward this loss are easy to spot. When the amount of effort required to retrieve information reaches a certain point, information is often set aside and eventually abandoned leaving it stored in formats that cannot be recovered. In time the contents of that information are forgotten. Now, it has been erased- a few more bytes for the memory hole. When this happens, we allow technology to disconnect the present from the past and sever us from our own roots (Marshall, John McClellan, 2020).

The Digital Preservation Coalition released its revised third edition of the 2021 BitList in November of that year. This is a list of "digitally endangered species" that offers a "snapshot of the concerns expressed by the global digital preservation community with respect to the risks faced by diverse types of digital content in varied conditions and contexts" (The BitList 2021: The Global List of Digitally Endangered Species, 2021). It is hoped that this list will assist digital preservationists to recognize the digital materials in their collections that are in the greatest need for urgent action to ensure the data remains viable. The BitList presents its information in groups by risk classifications ranging from Concern and Lower Risk to Practically Extinct. It also offers information by Digital Species. For the Portable Media species- floppy disks, DVDs, and USBs, to name a few- the general recommendation "for all portable storage media types is to plan and implement refreshment and replication as early as possible, moving the data to new forms of storage every 5-10 years" (The BitList 2021: Portable Media, 2021). However, it points out that the potential of quantifying the media in need is highly dependent on who is looking after it. Consider discs that accompany texts or magazines. In these situations, discs can be processed upon receipt. This in turn sets in motion the documentation of those media and where to find them allowing for a calculation of their storage needs on the library shelf. Yet, archives find themselves dealing with bit-level preservation of data coming from portable media and in legacy formats created in software such Shockwave Flash, Lotus 1-2-3, Word 1.0, and many more. With the method and tools but no time commitment, quality assessments will be few and far between. This makes calculating the necessary storage and resource commitments for this data an educated guess at best. Inevitably, there will be loss.

In the 2018 article Data, Data, Everywhere, Julie Engebretson pointed us toward the tidal wave of data being created, estimated that we add over 16 zettabytes (around 16 trillion gigabytes) to it each year, and reminded us that the end to this increase is nowhere in sight (Engebretson, 2018). Deciding what portion of this information holds historic value may come from various perspectives with varying recommendations, but we can be sure that some percentage does have such value. For now, how this portion of digital content will be processed- with activities that are siloed and disconnected or alongside and combined with its physical counterparts- will certainly be addressed on local levels, but format agnostic archival processing standards would be a step toward assured inclusion of digital assets into an institution's standard procedures and away from abandonment and loss of further pieces of our historical record. A record that, no matter what format or medium it comes in, should be shown the needed due diligence and respect that will ensure its availability for generations to come..

References

Appraisal. (n.d.). In Dictionary of Archives Terminology. Retrieved November 30, 2021, from dictionary.archivists.org/entry/appraisal.html

Arroyo-Ramirez, E., Dundon, K., & Peltzman, S. (2021, March 16). Works in progress webinar: Holistic approaches to born-digital appraisal and accessioning—Revising the UC guidelines for efficient archival processing. www.oclc.org/research/events/2021/031621-holistic-approaches-born-digital-appraisal-accessioning.html

Chen, S. (2001, March). The paradox of digital preservation. Computer, 34(3), 24–28.

Dundon, K., McPhee, L., Arroyo-Ramirez, E., Beiser, J., Dean, C., Eagle Yun, A., Jones, J., Liebhaber, Z., Macquarie, C., Michels, L., Peltzman, S., & Phillips, L. (2020). Guidelines for efficient archival processing in the University of California Libraries (Version 4). UCLA: Library. escholarship.org/uc/item/4b81g01z

Engebretson, J. (2018). Data, data, everywhere. Baylor Arts & Sciences, Fall 2018. blogs.baylor.edu/artsandsciences/2018/10/30/datascience

Kentucky Department for Libraries and Archives. (2002). Electronic records: Introduction. Commonwealth of Kentucky. kdla.ky.gov/records/recmgmtguidance/Pages/elecrecmgmt.aspx

Marshall, J.M. (2020). The modern memory hole: Cyberethics unchained. Athenaeum Review, Winter 2020(3), 94–101.

Preservation issues. (2015). In Digital Preservation Handbook (2nd ed.). Digital Preservation Coalition. www.dpconline.org/handbook/digital-preservation/preservation-issues

Preservation methods and techniques. (n.d.). In Dictionary of Archives Terminology. Retrieved November 30, 2021, from dictionary.archivists.org/entry/preservation-methods-and-techniques.html

The BitList 2021: Portable media. (2021). Digital Preservation Coalition. www.dpconline.org/digipres/champion-digital-preservation/bit-list/portable-media

The BitList 2021: The global list of digitally endangered species. (2021). Digital Preservation Coalition. www.dpconline.org/digipres/champion-digital-preservation/bit-list

The three essentials of digital preservation part 1: File storage. (2018). Sustainable Heritage Network. sustainableheritagenetwork.org/system/files/atoms/file/The_Three_Essentials_of_Digital_Preservation_Part1_File_Storage.pdf

Weber, C.S. (2020, April 28). Time estimation for processing born-digital collections. Hanging Together: The OCLC Research Blog. hangingtogether.org/time-estimation-for-processing-born-digital-collections