IJoL International Journal of Librarianship

ISSN: 2474-3542 Journal homepage: http://journal.calaijol.org

HathiTrust and Local Digital Stewardship: A Case Study in How Massive Digital Libraries Affect Local Digital Resources Decisions

Heidi M. Winkler and Joy M. Perrin

Abstract:

This article reviews the influence that massive digital libraries like the HathiTrust Digital Library can have on local, smaller institutions' digitization, preservation, and curation programs. The history of HathiTrust's digital preservation efforts as a Trusted Repository is reviewed. A case study is presented showing how one academic library made difficult digital stewardship decisions in a modern world of globally federated preservation initiatives. The authors introduce the concept of deselection as part of the digital curation process and discuss how digital collection administrators can refine their local digital preservation efforts to better reflect the realities of constrained human and financial resources.

To cite this article:

Winkler, M.H., & Perrin, M.J. (2017). HathiTrust and local digital stewardship: A case study in how massive digital libraries affect local digital resources decisions. *International Journal of Librarianship, 2*(1), 32-41. https://doi.org/10.23974/ijol.2017.vol2.1.11

INTERNATIONAL JOURNAL OF LIBRARIANSHIP, 2(1), 32-41
ISSN:2474-3542

HathiTrust and Local Digital Stewardship: A Case Study in How Massive Digital Libraries Affect Local Digital Resources Decisions

Heidi M. Winkler, Joy M. Perrin
Texas Tech University Libraries, Lubbock, TX, USA

ABSTRACT

This article reviews the influence that massive digital libraries like the HathiTrust Digital Library can have on local, smaller institutions' digitization, preservation, and curation programs. The history of HathiTrust's digital preservation efforts as a Trusted Repository is reviewed. A case study is presented showing how one academic library made difficult digital stewardship decisions in a modern world of globally federated preservation initiatives. The authors introduce the concept of deselection as part of the digital curation process and discuss how digital collection administrators can refine their local digital preservation efforts to better reflect the realities of constrained human and financial resources.

Keywords: digital weeding, digital curation, digital stewardship, levels of preservation, trusted repository, digital preservation

INTRODUCTION

Taycher (2010) estimated that there were nearly one hundred thirty million books published worldwide. The HathiTrust Digital Library claimed in late 2016 a collection of at least 7,364,677 digitized book titles, a number that is constantly growing (HathiTrust Digital Library, n.d.-a). The HathiTrust Digital Library is a "digital preservation repository" that was founded in 2008 through a multi-university collaboration to house books scanned by Google, the Internet Archive, and Microsoft, in addition to books scanned by the various partner institutions (HathiTrust Digital Library, n.d.-b). The HathiTrust Digital Library represents a massive effort to curate and house these digitized books.

If the Taycher estimate was correct, this leaves millions of books worldwide still restricted to physical format that would benefit from digitization and ultimately digital preservation. Identifying these items, scanning them, and making sure they are properly protected for the future are new challenges in this increasingly digital age. As individual organizations look to digitize their holdings, they must first address a number of questions, including which items should be marked for digitization, and what should the stewards of these items do with them once they are digitized?

This article addresses the influence that massive digital libraries like HathiTrust have on local institutions' digitization, preservation, and curation programs and the choices made about which items are preserved at the highest levels of preservation and which ones are not. This discussion requires clarifying what digital preservation and curation, which is collectively known as digital stewardship, is in this context.

BACKGROUND

Digital Stewardship

The bound book can be regarded as a reasonably reliable format. Books sitting on shelves in a library are fairly stable. If left alone, they will degrade far more slowly than if they are frequently checked out and used by patrons. The opposite is true for digital files and formats. If a digital file is left alone, untouched, and unchecked, it will almost certainly corrupt. The United States Library of Congress defines digital preservation as "the active management of digital content over time to ensure ongoing access" (Library of Congress, n.d., para. 1). Digital preservation can encompass both born-digital content (materials that originate in a digital environment, such as on a computer) as well as content that was reformatted from a physical form to a digital form (ALCTS, 2008). In the context of this article, all discussions of digital preservation will be in reference to the files created when a book is digitized through scanning.

Digital preservation follows four simple tenets:

Tenet 1: Make multiple copies of the digital content, and make sure those copies are backed up (Standford University, n.d.);

Tenet 2: Save those copies in geographically distant locations;

Tenet 3: Check on the copies regularly to make sure they have not been altered in some way (The National Digital Stewardship Alliance, 2014), and;

Tenet 4: Check the format of the items to make sure they are still usable (Digital Curation Centre, n.d.).

Digital files are particularly vulnerable to corruption and destruction (Digital Preservation Coalition, 2017). Typically, digital files are stored on hard drives or servers, which can crash or fail. If the digital file is stored on two or three different media formats, there is a greater chance that if one of the storage media fails, the other copies will remain safe. Every copy of a digital file must be checked at regular intervals via checksums to ensure their data is not being corrupted. Those files must also be prepared for transfer to new media as their original physical media, hardware, software, or formatting become obsolete.

Digital preservation is first and foremost concerned with making sure that a file's bits survive for the foreseeable (and unforeseeable) future; however, people and institutions cannot rest on simple preservation of the bits of the electronic file alone (ALCTS, 2007). The technical aspects of preserving digital files are fairly straightforward compared to the complexities of developing policies for managing a digital archive to avoid unintentional human error. The technical infrastructure of a digital preservation system could appear to work perfectly, but if there are no policies in place for how often administrators should check their data, file corruption could go unnoticed. Administrators need to put processes into place that detail how to go about correcting

any mistakes they or others make while handling digital data. The more people there are within an organization involved in the digital preservation process, the more likely it will be that human error will occur.

Preservation resides within and alongside digital curation, which focuses on building and managing the value of digital objects (Lazorchak, 2011). Digital content administrators must curate their materials so that they are continually useful. Digital curation may take the form of reorganizing collections, adding metadata, or enhancing digital files to make them useful to current and future patrons and administrators. Digital stewardship is the term given to the lifecycle management approach that encompasses both digital curation and digital preservation (Lazorchak, 2011). It may seem that the more difficult and expensive aspect of converting books to digital formats is the actual scanning, but the cost of long-term stewardship will continue for as long as the digital file exists. As Lazorchak wrote, digital stewardship is an active process; an organization cannot simply post a file online to its website or repository and consider it permanently preserved. Rather, institutions must take steps to ensure that their digital content is secure.

The past two decades have seen the advent of a partnership approach to managing digital content. Institutions form agreements with one another to curate digital collections from a central location and pool their resources to preserve this content for years to come. One remarkable example of this community digital management is the HathiTrust Digital Library.

The HathiTrust Digital Library: Trusted Repository

HathiTrust's mission is to preserve and make accessible history's broad cultural and academic records. By November 2016, the library had amassed over 600 terabytes of data (HathiTrust Digital Library, n.d.-a). A humanities researcher exploring one title could potentially find dozens of different digitized editions of that title, all with different editors, different introductions, different text, and even unique marginalia. HathiTrust gives users access to a wide berth of digitized academic materials from over 100 partner institutions. For example, the Library of the Complutense of Madrid alone used HathiTrust to provide global users access to the roughly 120,000 out-of-copyright books from the 16th to 19th centuries that it had digitized (Magán, Palafox, Tardón, & Sanz, 2011). With this much academic research and cultural heritage potential under its stewardship, HathiTrust must take strategic, purposeful steps in its digital preservation efforts to ensure these materials are available for generations to come.

In March 2010, HathiTrust became a certified Trusted Repository by the Center for Research Libraries (HathiTrust Digital Library, 2011). The criteria for being certified as a Trusted Repository, meaning that the institution has been assessed for its reliability and commitment to long-term digital preservation, is extensive. Organizations must not only document how they meet minimum requirements, but must be able to show evidence of their compliance. HathiTrust's certification required an audit that covered three major areas: organizational infrastructure; digital object management; and technologies, technical infrastructure, and security. Within each of these three areas was more specific criteria. According to its own website, HathiTrust utilizes various strategies in its work "to ensure the long-term integrity of deposited materials" (HathiTrust Digital Library, n. d. c, para. 1). It validates all its digital content upon ingesting it into its system and checks the integrity of all of its files on a quarterly basis. The library also employs digital preservation community-approved standards and formats that are supported across a variety of platforms, and it is confident in its ability to adapt and migrate to new preservation formats over

time. What is important to note about the *Trustworthy Repositories Audit & Certification* (TRAC) document (OCLC & CRL, 2007) is that while technical requirements for digital preservation are laid out for institutions to consult and follow, they are far outnumbered by administrative, policy-based requirements. The designation requires organizations to have strict policies that reduce the chance of human error.

The task of preserving materials securely is a prohibitively expensive undertaking for any single organization. It becomes feasible, however, when many different organizations pool their resources. When an institution joins the membership of HathiTrust, it is contributing to the long-term preservation of all the volumes that are held in this massive digital library. In the following case study, one university library realized the positive effect that acting in concert with HathiTrust's holdings could have on its own digital stewardship decisions.

DIGITAL STEWARDSHIP AT THE INSTITUTIONAL LEVEL

When the Texas Tech University (TTU) Libraries began digitizing content in 2004, digital preservation was an afterthought. Like many organizations excited about digitization, the project teams creating digital collections were focused on access. Getting digital content online as quickly as possible outweighed any concerns about its long-term survival. The library primarily digitized out-of-copyright books. The scanning effort ramped up just as Google announced its massive book digitization project (Google, n.d.). As the Google Book project picked up speed, interest and support of book digitization by the library waned on the assumption that all the books had been scanned by other organizations. The staff involved did not realize that some of the items they had scanned were not available elsewhere and were in need of additional preservation.

The library was also involved in a nascent electronic theses and dissertations (ETD) initiative. It did not take long for the ETD collection database to twice experience corruption (Perrin, Winkler, & Yang, 2015). At the time, backups done nightly and then retained for 30 days were considered sufficient protection against any malfunctions. In both cases, however, the corruption was not noticed until well after the 30-day window for the backup had passed. While the collection's PDF (Portable Document Format) files were undamaged, a year's worth of metadata was lost and had to be redone. These incidents taught the library two vital lessons about digital preservation. First, digital content is more than just files. Effective digital library content is a combination of both an item's file and its metadata, and both must be protected equally. Second, 30-day backups only work if it can be guaranteed that the items will be checked at least every 30 days.

These crises led the library to invest in basic digital preservation. It acquired a second storage site located miles from the university campus in order to maintain geographically distant copies, providing a small level of protection. Updates to the digital archive on campus are replicated immediately to the remote site. Physical access to the offsite storage continues to be restricted to a few information technology (IT) staff members in order to keep human error to a minimum. Checksums are regularly confirmed on both the files on campus and the off-site storage to protect against file corruption. All files that are stored in both locations are considered to be in the dark archive.

The digitized collections benefitted from the digital preservation measures taken to protect the born-digital collection. In addition, the library's Digital Resources Unit (DRU) had at the time

a policy to save every single file created during the digitization process just in case. Since the team had no guidance from content owners about what was important, everything was kept, including failed projects or items that were scanned but never intended to be online. All this digital hoarding resulted in the dark archive server ballooning to 10 terabytes in both of its locations (20 terabytes total). Furthermore, these files were not organized consistently. Each project contained different file formats and utilized different naming schemas. Some collections even contained duplicated files from multiple scans. While the files themselves were secure, the DRU realized that the collections needed curation.

In 2013, the unit's new Digital Stewardship Librarian (DSL) was given authority over the library's dark archive servers and the responsibility of managing the digital preservation and curation efforts of the library. This librarian was tasked with organizing the digital archive, bringing the older collections already in the dark archive up to newer curation standards, and preparing new content for preservation. She began by reviewing the content in the dark archive and assessing each collection for its individual digital preservation needs. She had to approach the archive from a different perspective than the "preserve everything just in case" model. That process began with gaining context about why and how each digital collection had come into being over the previous decade. The digital collections had begun with the intent to digitize, in the name of open access, as many out-of-copyright books from the library stacks as possible. Over time, the DRU had refined its approach to focus on the preservation of materials unique to the university, such as ETDs, university yearbooks, and other special collections. Now the DSL had to make decisions about collections that no longer met the mission of the digital library and were taking up space that could be used for other materials.

The DSL selected seven digital collections that had been scanned from textbooks, cookbooks, copies of state legislative records, novels, and plays that were no longer within the mandate of the dark archive. She identified HathiTrust's catalog as a tool against which she could compare these collections, since it was also a repository of scanned book content, albeit much larger. She decided to evaluate exactly how much overlap existed between these seven collections and HathiTrust. The DSL's evaluation addressed replicated content as opposed to replicated editions. Since this librarian was the only one assigned to the project, she did not have the time to individually investigate what could often be dozens of editions of the same book in the HathiTrust catalog. Rather than getting into the minutiae of what to do if the library had the 1892 edition of a textbook and HathiTrust had the 1891 edition, for example, she tried to gauge quickly whether the contents of the books were essentially identical. When they were, she marked the TTU copies as part of the overlap.

The DSL discovered that of the 391 individual records comprising the seven scanned collections, 307 were also being preserved as part of the HathiTrust collections. The DSL decided to weed these 307 items from the local digital archive. A process for how to record which items were deselected and why needed to be created before the items were removed irrevocably. In the dark archive, the librarian started keeping a log of archive changes and a list of policies. The log kept records that stated which items had been moved or removed from the dark archive and why. If someone looked for an item available online and could not find it in the archive, they would have an explanation for why it was not there. The librarian created a policy stating that items would sit in "Ready to Delete" for two weeks as project stakeholders were notified of their removal from the dark archive and the reasoning behind it.

While there was significant overlap between the TTU collections and HathiTrust's holdings, the librarian also identified items that Texas Tech had that HathiTrust did not have. Those items would be moved into a second group to eventually be submitted to the HathiTrust collection and removed from the library's dark archive.

The library decided, however, that public online access to the removed content would continue to be maintained through its DSpace, which is hosted off-campus through a statewide consortium. It does not cost the university extra to continue offering access to these items. From the patron's perspective, nothing has changed. Instead, preservation copies have been carefully deselected from the locally maintained dark archive in order to free up resources to apply digital preservation measures to items that are unique to the university. Since public access to the materials would continue, the DSL did not take account of usage statistics or the age of the items in her assessment. Instead, she focused on rarity and determining whether the library held the only copies of the content. The collections' sponsors approved the items being removed from the dark archive thanks to continued public access.

The Levels of Preservation

The library has since adopted a policy to assign collections a level of preservation that dictates how items will ultimately be processed into the dark archive. This policy ensures that everyone involved in the workflow understands the importance of the collection and can handle items properly without wasting resources on lower-level collections.

The library assigns a low Level One status to its online-only digital content. The items are objects that are not unique to the university that the library has confirmed are being preserved in a Trusted Repository like the HathiTrust. Most of the items that have been assigned to this level are part of the library's legacy collections and would likely be rejected if proposed now. Since, however, the repository's web analytics do indicate that these copies get some use, their online presence will be maintained. The items removed from the archive in the deselection project described above constitute Level One items. When books are brought in for digitization, the DRU first checks HathiTrust to determine if the book already exists in some form in its catalog. If the item is in HathiTrust, the book is not scanned and the item never even achieves a Level One Status.

Level Two items are born digital; namely, the theses and dissertations that have been submitted electronically to the university since 2005. In this level, the display copy PDF is the only copy the library publishes online and preserves in the dark archive.

Level Three encompasses items that are physically unique to the university and have been scanned. In this case, the display copy is online, while both the archive-quality TIFF (Tagged Image File Format, .tif) files and display-quality JPEGs (Joint Photographic Experts Group, .jpg) or PDFs are in the local dark archive. Items are scanned as a 600 DPI (dots per inch) full color TIFF, but in order to display them effectively online, the quality is reduced for the display copy. In some cases, the online copy has a watermark. The archive copies contain more information than the display copies, and so the extra TIFF files are kept. Saving these files is especially important for cultural heritage materials.

Level Four is for items that cannot be displayed online but are unique to the university and born digital. Embargoed ETDs are kept offline temporarily or permanently to respect publication

holds, grant obligations, or occasionally, author safety. Because of this, the official copies of these items only exist on the dark archive server.

DISCUSSION AND FUTURE RESEARCH DIRECTIONS

The authors noted throughout this process that the priorities that guide deselection within traditional physical collections do not work in the same way for digital weeding processes. For a traditional physical collection, heavy usage by patrons signals that an item has earned its place in the stacks and needs extra protection and care. For a digital item, though, heavy usage is not necessarily as important of a measure as is its rarity online. Some digital content can be culturally significant but not necessarily unique. Take, for example, the *Legislative Manual* for the 37th Legislature of the State of Texas. Any student of Texas state legislative history would be able to use the content of this document to pore over legislature rules as well as the rolls of standing committees for the 1921 state legislature. These students would easily find this manual and others like it in the HathiTrust catalog, but they may be surprised to find that these documents have been digitized many times by many other institutions. In the grand scheme of digitization, it does not cost much to scan one item and make it available online, but if all of these organizations apply the same levels of digital preservation to their respective copies of this single item, it would be a waste of resources. Instead, if the item is not already in HathiTrust, one of them could submit it to ensure that it is preserved. Each institution could subsequently treat the item like a Level One item.

What makes an item unique? If an institution has the only official copy of a work, or if its copy is a rare edition or has noteworthy marginalia, it may be considered unique. For all practical purposes, a unique book would be one that is not in HathiTrust. Professionals involved with local digitization may not realize that some of their unique items have cultural significance and are worthy of greater preservation efforts. It would also be a mistake to assume an entire collection needs to be preserved locally if an organization like HathiTrust, with greater resources and experience, can house some or all of the collection.

The authors recognize that digital deselection may not be a comfortable process for librarians. After all, traditional physical book weeding has long been considered one of the least desirable library tasks (Snyder, 2014). Many librarians feel uncomfortable taking on a project that entails permanently removing materials from the library; Slote (1975) noted that people have an almost "sacred" regard for books and that their deselection from a collection can become "painful" (p. 5). When one considers the emotional connections librarians can have with physical collections, then it is not surprising that they are equally, if not more, defensive of their digitized book collections. At TTU alone, some of the library's digital book collections have taken years to complete and are as a result technically worth tens of thousands of dollars. With such a sunken cost mentality associated with digitization, the realization that a collection might not be worth local digital preservation can be difficult to process, but it is necessary if the collection's value to the library does not justify the expense.

While HathiTrust currently only handles books, a similar model can be adopted for other formats. The Minnesota Digital Library Image Preservation Prototype Project, for example, created a prototype for how to deposit images into HathiTrust for "access, storage, and preservation purposes" (HathiTrust Digital Library, n.d.-d, para. 1). The project found certain problems with expanding the infrastructure of HathiTrust to other formats besides books. One

major difficulty that the project stumbled upon was that organizations often do not keep their master images formatted to HathiTrust's specifications (Celeste & Skinner, 2010). They also realized that institutions uploading images to HathiTrust might experience difficulty mapping their local metadata to the schema that HathiTrust requires. The project's findings indicate that what is good for digital book preservation may not work well when expanded to other types of digital preservation. Additionally, rights issues with images and other media may be more complex than with books, and institutions may feel less comfortable signing away rights to their image collections. Instead of trying to make HathiTrust work for other formats, academia might find it more beneficial to their content to create separate digital library systems for different digitized formats such as images, video, and audio. These systems would allow each format's unique metadata and formatting problems to be handled while still being supported through a consortial model to pay for the curation and preservation of the items.

CONCLUSION

Digital stewardship concerns the preservation and curation of digital objects over the course of their lifecycles. These efforts have traditionally been internal institutional matters; ensuring the survival of digital files was a local issue that would be dealt with in different ways and with different budgets. Not all digital preservation efforts are created equal, however, not all digital content needs to be given the same consideration for local preservation. As each organization plans their digital projects, a part of every project should include an evaluation of the collection's unique stewardship needs. In the new global partnership approach to digital preservation and curation, smaller institutions must consider the limits of their preservation and curation resources. They must decide what they must absolutely steward themselves to do, and what would be better stewarded by a trusted repository with more extensive resources.

Local theses and dissertations, yearbooks, or other unique community publications may be more valuable items to digitize than other items that are considered more traditionally culturally significant. Not all organizations will choose to handle their items the same way, but no matter how they choose to move forward, they need to take into consideration the greater digital library world. Instead of each institution expending funds and efforts to preserve items, it might be best to funnel the items and funds to a central organization like HathiTrust so that they can be preserved more efficiently and at a higher level.

References

ALCTS Preservation and Reformatting Section, Working Group on Defining Digital Preservation. (2007, June 24). *Definitions of digital preservation*. Retrieved February 18, 2015, from http://www.ala.org/alcts/resources/preserv/defdigpres0408

Celeste E. & Skinner K. (2010). *Minnesota Digital Library and HathiTrust image preservation prototype project report: Executive summary.* Minnesota Digital Library. Retrieved June 30, 2017 from http://www.mndigital.org/projects/preservation/esReport.pdf

Digital Curation Centre (n.d.) *What is digital curation?* Retrieved July 6, 2017, from http://www.dcc.ac.uk/digital-curation/what-digital-curation

Digital Preservation Coalition (2017). *Why digital preservation matters. Digital preservation handbook*. Retrieved July 6, 2017, from http://www.dpconline.org/handbook/digital-preservation/why-digital-preservation-matters

Google. (n.d.). *Google books history*. Retrieved June 30, 2017 from https://www.google.com/googlebooks/about/history.html

HathiTrust Digital Library. (2011). *HathiTrust trustworthy repository audit and certification (TRAC).* Retrieved June 30, 2017 from Hathi Trust Digital Library:from https://www.hathitrust.org/trac

HathiTrust Digital Library. (n.d.-a). *Welcome to HathiTrust!* Retrieved June 30, 2017 from https://www.hathitrust.org/about

HathiTrust Digital Library. (n.d.-b). *Our partnership*. Retrieved June 30, 2017 from https://www.hathitrust.org/partnership

HathiTrust Digital Library. (n.d.-c). *Digital preservation policy*. Retrieved June 30, 2017 from https://www.hathitrust.org/preservation

HathiTrust Digital Library. (n.d.-d). *Minnesota digital library image preservation prototype project*. Retrieved June 30, 2017 from https://www.hathitrust.org/mdl_images

Lazorchak, B. (2011, August 23). *Digital preservation, digital curation, digital stewardship: What's in (some) names?* Retrieved June 30, 2017 from https://blogs.loc.gov/thesignal/2011/08/digital-preservation-digital-curation-digital-stewardship-what%E2%80%99s-in-some-names/

Library of Congress. (n.d.). *Digital preservation: About*. Retrieved July 3, 2017 from http://www.digitalpreservation.gov/about/

Stanford University (n.d.) .*Preservation principles.* Retrieved July 6, 2017, from https://www.lockss.org/about/principles/#decentralized

Magán, J. A., Palafox, M., Tardón, E., & Sanz, A. (2011). Mass digitization at the Complutense University Library: Access to and preservation of its cultural heritage. *LIBER Quarterly, 21*(1), 48–68. DOI: http://doi.org/10.18352/lq.8007

The National Digital Stewardship Alliance . (2014) *Checking your digital content*. Retrieved July 6, 2017, from http://ndsa.org/documents/NDSA-Fixity-Guidance-Report-final100214.pdf

OCLC, & CRL (2007). *Trusted repositories audit & certification: Criteria and checklist.* Chicago: CRL, The Center for Research Libraries. Retrieved June 30, 2017, from http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf

Perrin, J., Winkler, H., & Yang, L. (2015). Digital preservation challenges with an ETD collection — A case study at Texas Tech University. *The Journal of Academic Librarianship, 41*(1), 98-104. doi:10.1016/j.acalib.2014.11.002

Snyder, C. E. (2014). Data-driven deselection: Multiple point data using a decision support tool in an academic library. *Collection Management, 39*(1), 17-31. doi:10.1080/01462679.2013.866607

Taycher, L. (2010, August 5). *Books of the world, stand up and be counted! All 129,864,880 of you.* Retrieved June 30, 2017 from http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html

---

About the authors

Joy M. Perrin (ORCiD: 0000-0001-5524-9071) is the Digital Resources Librarian at the Texas Tech University Libraries. She holds a Master of Library Science from the University of North Texas. Ms. Perrin has ten years' experience with digital projects and is the author of the 2015 book Digitizing Flat Media: Principles and Practices.

Heidi Winkler (ORCiD: 0000-0003-4645-9741) is the Digital Stewardship Librarian at the Texas Tech University Libraries. She holds a Master of Science in Information Studies from the University of Texas at Austin. Ms. Winkler's research interest include digital preservation and curation, and her work has appeared in a number of peer-reviewed journals.