Difficulties with Digitization

R. Sumi Matsumoto

FA-INFO-601-06

Assignment 3

Review of an issue/trend in the information field

21 November 2019

The Introduction

Libraries have been engaged in digitization projects since the 1990s, but the popularity of digitization in libraries and museums continues to increase (Lopatin, 2006). A 2015 survey of libraries and museums currently engaged in digitization projects found that over 90% believed time spent on digitization would stay the same or increase over the next year and some institution even projected digitizing up to 5,000 objects per year (Primary Research Group, 2015).

Cultural institutions spend time and money on digitization projects because digital surrogates offer some benefits over physical objects (Siggins, 2008). Digitization increases access by allowing multiple users to view an object simultaneously, an object to be viewed from any distance, and fragile documents to be viewed by the public (Karlsson, 2019; Lopatin, 2006). Digital assets can be re-contextualized and placed in multiple exhibits and collections (Bülow, Spencer, & Ahmon, 2011). These digital exhibits can also be used to raise the profile of an institution outside of its local area (Lopatin, 2006). Digital assets also reduce demand for and wear on original objects (Karlsson, 2019). Digitization projects may increase a repository's physical and intellectual control of their collections since digitized objects require item-level description where most repositories currently only have collection-level descriptions (Proffitt, 2015).

Digitization interests me because I believe it is an intersection of my aspirations to work with special collections and my background in computer science and database management. However, digitization is a relatively new preservation technique, and there are many debates about its benefits versus dangers (Conway, 2010). As a self-proclaimed "millennial luddite," balancing enthusiasm and skepticism is something I already apply to technological innovations. The benefits of digitization are often touted, so I wanted to dig deeper into the major challenges

with digitization projects. In this paper, I will outline the difficulties with digitization projects, focusing on three main problem areas: project costs, copyright claims, and the future of digital assets. For each problem area I will explain the challenge presented as well as survey the various, imperfect solutions currently in practice.

Project Costs & Resource Inefficiency

Digitization is resource-heavy (Bülow et al., 2011). A survey of 61 international libraries and museums involved in digitization projects found that institutions spend on average about $163,000 annually on digitization, with the highest budget over $2,000,000 (Primary Research Group, 2015). However, a budget does not accurately reflect the actual costs needed to complete a project. Budget often reflects an institution's interest in digitization and the endowment of the institution, as over 50% of digitization projects are funded primarily by the institution (Lopatin, 2006). This money is typically not being fundraised; it is being reallocated from other parts of the budget. Most often, these cuts are in the preservation budget (Conway, 2010). This means that while digitization is touted to help preserve the original object, it may endanger long-term physical preservation if a lack of funds disallows proper preservation conditions and care.

Two main alternatives to internal funding are grant support and commercial partnership (Bülow et al., 2011). Grant support for digitization projects often comes from large organizations like the Institute of Museum and Library Services or the Andrew W. Mellon Foundation (Conway, 2010; Lopatin, 2006). Grants have restrictions on who can apply and may require dollar-for-dollar matching funds. Grant-funded projects also have requirements that must be followed: including a timeline, restrictions on funding allocation, and pre-established deliverables (STLPG Division National Park Service, 2019). Commercial partnerships allow

cultural institutions to outsource digitization projects for no to nominal cost, however these partnerships also allow the institution less control over the digitization project (Bülow et al., 2011). The commercial partner is responsible for scanning and file transmission while the library or museum supplies the collection and receives a copy of the digital files created (University of Michigan Library, 2005). However, collections chosen to be digitized are limited to the interest of the commercial partner: archives' biographical collections have value to genealogical organizations like Ancestry.com while libraries' book collections interest organizations like ProQuest or Google (Bülow et al., 2011; Karlsson, 2019). This overlooks the digitization need of the objects—objects most likely to degrade—and the institution—objects that are pulled most frequently (Proffitt, 2015). The commercial partner may also commodify the digitized works, selling access to full versions of an object (Karlsson, 2019; Siggins, 2008).

Commercial partners and grant sponsors are both fixed-term funding solutions for digitization projects. Collections chosen for digitization are restricted by these external organizations, and funding typically only covers the scanning process. This means these solutions do not account for the cost of continuing storage or care of these digital assets. Additionally, only grants that include fund allocations for hardware purchases and staff training build internal infrastructure for future digitization projects (Conway, 2010). Commercial partnerships are effective for large-scale digitization projects that would be impractically cost, time, and resource draining on an institution (University of Michigan Library, 2005). Grant funded digitization projects are effective for digitizing specific collections with a time-sensitive need. The ideal solution would be for institutions to pursue endowments earmarked for digitization projects, as it would create a long-term, steady, internal source of money, which could be used based primarily on the need of the institution and its collections. Currently only a

small percentage of libraries and museums in the United States of America receive funding for digitization projects this way (Primary Research Group, 2015).

In addition to monetary costs, digitization projects require heavy use of personnel. On average, four staff members per institution spend part of their day on a digitization project: averaging about 5,400 hours of total work annually (Primary Research Group, 2015). Digitization projects require staff from many different disciplines including curation, conservation, legal, IT, cataloguing, and marketing (Bülow et al., 2011; National Archives, 2016). Because digitization projects require inter-department collaboration, it is crucial for an institution to have support from upper management and administration as well (Primary Research Group, 2015). Currently, most preservation specialists do not have the technical skillset required for digitization projects, and imaging specialists do not have object-handling knowledge (Conway, 2010). In addition to imagining, a digitization project requires adding metadata to each digital asset; without this metadata, digital assets are difficult to find and any ease of access gained by having digital surrogates is counteracted (Siggins, 2008). However, most cultural institutions only have collection-level description so this item-level metadata needs to be created (Proffitt, 2015). Nearly 50% of staff time on digitization projects was spent on cataloguing and metadata issues (Primary Research Group, 2015).

In order to avoid overextending current staff, some digitization projects budget for project staff (Bülow et al., 2011; National Archives, 2016). This is an effective solution if there is the budget to support it. Although it does not help fill the current need for professionals with a cross-discipline skillset in technology and preservation, programs and courses are emerging to bridge this knowledge gap in aspiring GLAM professionals (Conway, 2010). Alternatively, there are services that specialize in the digitization of archival and special collections materials to

which digitization projects can be outsourced (National Archives, 2016). I foresee digitization and digital asset management continuing to be an important aspect of GLAM professions, so I believe pursuing longer-term solutions like investing in training current staff and fostering an inter-disciplinary skillset in aspiring professionals is the best solution to reducing the number of staff and departments required to work in tandem on a digitization project. Outsourcing can be a useful alternative for short-term projects, like digitizing items in immediate danger of degradation. To help with metadata efforts, some institutions utilize crowdsourcing. Crowdsourcing most often uses the public—contacted through an institution's social media—to help identify unknown, or mystery, objects (Primary Research Group, 2015). A growing proposition is to crowdsource metadata from scholars and researchers, who are the largest user-base for digitized collections (Bülow et al., 2011; Proffitt, 2015). This allows the type of information users will search for to be included in the metadata scheme. It is also not necessary to add full metadata to every item; some collections can be adequately searched with aggregate metadata and OCR scans (Proffitt, 2015).

Copyright Claims & Ownership

Museums and archives already face legal and ethical issues over the ownership of physical objects, and digitization creates a new set of legal dilemmas surrounding ownership and copyright (Karlsson, 2019; "Ownership disputes," n.d.). An institution's lack of ownership over the copyright of an object may prohibit the item from being digitized, creating holes in the digital version of a collection (Callaghan, 2017). This inconvenience is problematic because it leads to gaps in public knowledge; digitization efforts are already creating bias in cultural heritage due to the dissemination of only information housed by countries and institutions with the infrastructure

to undertake mass digitization projects (Siggins, 2008; United States Copyright Office, 2015). In addition to having proper ownership to reproduce an object digitally, institutions that display digital collections in public, online catalogues must protect their rights to these digital assets. Oppositely, while the institution physically controls public domain objects in its collection, digital surrogates are open to any replication. These public digital surrogates allow for use of the object by individuals or institutions who, unlike the library, museum, or archive, are not motivated by the object's conservation (Karlsson, 2019).

When an institution does not own the copyright to an object in its collection, it can get permission from the owner in order to digitize the object. Although this is seemingly straightforward, the process involves identifying the copyright holder, contacting the copyright holder, clearing the use of the copyrighted material, specifying use and further reproduction restrictions, and getting supporting documentation (Lopatin, 2006). This process can be further complicated when the rights of ownership to the original object are being disputed (Karlsson, 2019). Libraries and archives can take advantage of a legal exception that allows for up to three copies of a copyrighted work to be made by the institution; these copies can be digital, but they can not be shared publically (Besek, 2003). Libraries can also cite fair use law as a way to get around copyright: claiming digitized items are not-for-profit and used educationally (Lopatin, 2006). However, fair use law is not explicit in its precedent, and can be a risky claim to make (Besek, 2003). In an effort to protect public digital assets from further reproduction, use statements should accompany all published digital assets. Unfortunately, this is a low threshold of security against piracy (Lopatin, 2006). Some institutions will claim copyright or exclusive reproduction rights on all digital surrogates created, even if the original objects fall into public domain. The legal precedents for this course of action are database rights and anthology rights,

where reproduction rights and copyright is granted on the basis of an institution's ownership of the collection as a whole rather than an individual object's provenance (Karlsson, 2019). An alternative solution is to only publically share copies of a digital asset with a watermark or with resolution low enough to disallow effective printing or publishing (Lopatin, 2006).

I do not believe that legally risky solutions are reliable solutions. Although potentially effective, they do not build strong precedent in order to create a legally sound, widely applicable, long-term solution for all institutions. These options are most viable for larger institutions that are able to afford to take greater legal risks due to in-house legal teams and more flexible budgets (Callaghan, 2017). Most objects in a collection can legally be digitized for preservation even if they can not be shared in a public online catalogue. This means there may be gaps in an accessible version of the digital collection, but the institution would have a complete digital copy. However, digitization is most commonly undertaken as a method of increased accessibility—rather than as strictly a preservation method—so seeking ways to make these digital surrogates publically accessible is important. Currently, I believe the best course of action is to ask for permission from the copyright holder, which can be costly but has been shown to be effective (Lopatin, 2006). Internationally, extended collective licensing schemes have been used to allow the mass digitization of copyrighted collections within specific scopes, including non-profit educational and research purposes. This scheme has been considered by the United States of America and if put into effect would greatly increase a cultural institution's legal protection against copyright litigation (United States Copyright Office, 2015). However there is no clear licensing precedent for orphan works in the United States, and international efforts, like EU IPO, are too cost-prohibitive for most cultural institutions (Callaghan, 2017). Orphan work legislation was considered in both 2006 and 2008, but no legislation was adopted.

Until there is specific legislation for mass digitization and orphan works, the best practice for cultural institutions is to claim fair use (United States Copyright Office, 2015). Although claiming fair use is not an ideal course of action, I believe cultural institutions should mass digitize and digitize orphan works to prove the need for clearer legislation on these issues.

Technological Advancement

Digitization is a relatively new preservation technique; microfilm, a comparable preservation technique of copying original objects, was created in the 1800s and has been used to keep copies of records for about 100 years—longer than modern computers have existed (Lopatin, 2006; "The History of Microfilm: 1839 To The Present," n.d.). Because digitization is still young, the long-term viability of digital assets is unknown. An important part of digitization is the output file format. This effects not only the quality of the file, but also the robustness of the file (Rieger, 2016). Due to technology's rapid change, there are already obsolete file formats that make accessing some digital data difficult if not impossible. Additionally, the medium on which digital information is stored is not robust; drives become demagnetized over time (*Vint Cerf on the prospect of a "digital dark age,"* n.d.). Higher quality output files are able to capture more information from the original object, but high-resolution files are large and take up a lot of memory. Digital memory also takes up physical space: even cloud storage (Bülow et al., 2011). Currently, many cultural institutions do not have policies or standards for digital asset management that are as thorough as those for physical objects (Conway, 2010).

Most digital access issues are with digital-born materials rather than digitized objects, because digital assets are being created with modern, readable file formats (*Vint Cerf on the prospect of a "digital dark age,"* n.d.). Looking forward, software often has backward

compatibility built in, which gives the user time to export unsupported files into a readable format. Although file reformatting may be time consuming and lead to some loss of information, I believe it is a better solution than fully re-digitizing materials because it is less expensive and will preserve the original object. Storing information is cheaper than ever (Mearian, 2017). Information can be stored and backed up on the cloud by third-party services safely and inexpensively. Although an institution loses some control by storing their digital collection with a third party, most cultural institutions do not have the resources to have a secure data storage facility. With the introduction of cloud-based storage solutions, degrading hardware is less of a concern for digitization projects (Bülow et al., 2011). There are currently no standards in place for digital asset management, but cultural institutions should create internal policies. These policies should be similar to those in place for physical objects as well as align with the purpose of the digitization project (Federal Agencies Digitization Guidelines Initiative-Still Image Working Group, 2009).

Unfortunately, reformatting files will potentially not be a viable solution moving forward. Cutting edge technology is pushing quantum computers (Baumhof, 2019). Quantum computing relies on a fundamental change in how information is stored and processed: from binary bits to quantum bits, also called qubits. This structural change would make all current file formats, hardware, and software obsolete (Rieffel, Polak, Gropp, & Lusk, 2011). I use the example of quantum computing because it is an extreme, yet realistic, example of the large leaps in technology, which exacerbate the fragility of digital collections. However rather than completely writing off digital mediums as a form of preservation, I believe that cultural institutions should work to stay on the cutting edge of technology. Being aware of new hardware and software that can improve digitization and digital preservation efforts will create the most robust digital assets

in the future. Additionally, if cultural institutions invest interest in technology, then digitization and digital preservation will become part of the conversation about cutting edge technologies (Langston, 2019).

The Conclusion

I believe that with careful planning, digitization can be an effective preservation tool for cultural institutions. Based on my analysis of the problems and effectiveness of the current solutions, I propose an ideal digitization project that minimizes difficulty within the limitations of today's technology and resources. Fundraising and the creation of an endowment will ensure long-term funding for both digitization projects as well as digital asset management and care without impacting the budget for or quality of physical object care. Grant funding can be used to outsource time-sensitive projects. Hiring archival digitization and digital preservation specialists will avoid spreading current staff too thin and ensure specialized knowledge. Metadata standards and schemas, digital asset management policies, and future-proofing techniques should be outlined from the onset and can be used for all digitization projects undertaken by the institution. These systems must be at least as effective and thorough as those in place for physical collections. The cultural institution should stay on top of new digitization technologies, cloud storage options, and digital asset management techniques, adapting with the changing climate of digital preservation.

Even the ideal solution proposed is imperfect. In any digitization project, original objects are prone to damage through exposure to light and heat as well as handling and transportation. These damages are mostly negligible or are the byproduct of any object handling regardless of use (Bülow et al., 2011). The ideal solution is also not practical for most cultural institutions. It

is viable and necessary for all institutions undergoing digitization projects to create policies for digitization and digital assets. All institutions should pursue long-term funding options, but immediate projects can utilize different short-term funding options based on the needs of the project. Copyright is a very limiting factor in digitization projects, and the course of action chosen should be at the discretion of the institution based on the institution's ability to manage risk. However, it is in the best interest of cultural institutions to push for clearer copyright legislation on mass digitization and orphan works; by pushing ahead with projects involving these issues, institutions are proving the need for this kind of legislation. I believe achieving these copyright policies is possible and an important part of making digitization projects feasible for smaller cultural institutions.

The future of digitization and digital asset management is unknown, but demand for digital access is increasing (Siggins, 2008). In order to stay relevant, cultural institutions must work toward digitizing their collections. In general, taking long-term approaches to problems are the best course of action. Unfortunately, most cultural institutions do not have the resources to do this. Even though the process is imperfect and unstandardized, digitization is an important investment for cultural institutions to make.

Sources

Baumhof, A. (2019, October 29). Quantum computers: Why Google, NASA and others are putting their chips on these dream machines. Retrieved November 18, 2019, from World Economic Forum website: https://www.weforum.org/agenda/2019/10/quantum-computers-next-frontier-classical-google-ibm-nasa-supremacy/

Besek, J. M. (2003). *Copyright Issues Relevant to the Creation of a Digital Archive: A Preliminary Assessment*. Council on Library and Information Resources and the Library of Congress.

Bülow, A., Spencer, R., & Ahmon, J. (2011). *Preparing Collections for Digitization*. Retrieved from http://eds.b.ebscohost.com.ezproxy.pratt.edu:2048/eds/ebookviewer/ebook/bmxlYmtfXzU1ODQ1Ml9fQU41?sid=12cf0ba1-d76b-48c1-ab03-7f6664b368ad@sdc-v-sessmgr01&vid=9&format=EB&rid=3

Callaghan, S. (2017). Has the introduction of orphan works licensing schemes solved the problem that orphan works present to digitization projects? *Archives and Records*, *38*(2), 244–256. https://doi.org/10.1080/23257962.2017.1327847

Conway, P. (2010). Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *The Library Quarterly: Information, Community, Policy*, *80*(1), 61–79. https://doi.org/10.1086/648463

Federal Agencies Digitization Guidelines Initiative-Still Image Working Group. (2009, November 4). *Digitization Activities*. Retrieved from http://digitizationguidelines.gov/guidelines/digitize-planning.html

Karlsson, Á. G. (2019). Copyrighting the Copies. *The Library Quarterly*, *89*(4), 333–347. https://doi.org/10.1086/704961

Langston, J. (2019, November 4). Project Silica proof of concept stores Warner Bros. 'Superman' movie on quartz glass. Retrieved November 19, 2019, from Innovation Stories website: https://news.microsoft.com/innovation-stories/ignite-project-silica-superman/

Lopatin, L. (2006). Library digitization projects, issues and guidelines. *Library Hi Tech*, *24*(2), 273–289. https://doi.org/10.1108/07378830610669637

Mearian, L. (2017, March 23). CW@50: Data storage goes from $1M to 2 cents per gigabyte (+video). Retrieved November 19, 2019, from Computerworld website: https://www.computerworld.com/article/3182207/cw50-data-storage-goes-from-1m-to-2-cents-per-gigabyte.html

National Archives. (2016, August 15). Strategy for Digitizing Archival Materials. Retrieved November 17, 2019, from National Archives website: https://www.archives.gov/digitization/strategy.html

Ownership disputes. (n.d.). Retrieved November 17, 2019, from Museums Association website: https://www.museumsassociation.org/ethics/27012016-ethics-case-studies-ownership-disputes

Primary Research Group. (2015). *Survey of Library & Museum Digitization Projects*. Retrieved from http://eds.b.ebscohost.com.ezproxy.pratt.edu:2048/eds/ebookviewer/ebook/bmxlYmtfXzE0MTA0NTBfX0FO0?sid=12cf0ba1-d76b-48c1-ab03-7f6664b368ad@sdc-v-sessmgr01&vid=11&format=EB&rid=14

Proffitt, M. (2015, March 23). Digitization challenges—A discussion in progress. Retrieved November 17, 2019, from Hanging Together: The OCLC Research blog website: http://hangingtogether.org/?p=4945

Rieffel, E. G., Polak, W. H., Gropp, W., & Lusk, E. (2011). *Quantum Computing: A Gentle Introduction*. Retrieved from http://ebookcentral.proquest.com/lib/pratt/detail.action?docID=3339229

Rieger, T. (Ed.). (2016, September). *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files*. Retrieved from http://digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf

Siggins, J. A. (2008). Unintended Consequences: A Friendly User Looks at User-Friendly Digitization. *Journal of Library Administration*, *48*(2), 195–206. https://doi.org/10.1080/01930820802231393

STLPG Division National Park Service. (2019, October 2). *SAT Preservation Notice of Funding Opportunity P20AS00005*. National Park Service.

The History of Microfilm: 1839 To The Present. (n.d.). Retrieved November 17, 2019, from Southern Regional Library Facility website: http://www.srlf.ucla.edu/exhibit/text/BriefHistory.htm

United States Copyright Office. (2015). *Orphan Works and Mass Digitization: A Report of the Register of Copyrights* (p. 234).

University of Michigan Library. (2005, August). *UM Library/Google Digitization Partnership FAQ*. The University of Michigan.

Vint Cerf on the prospect of a "digital dark age." (n.d.). Retrieved from https://www.cbsnews.com/video/vint-cerf-on-the-prospect-of-a-digital-dark-age/