SJSU | SCHOOL OF INFORMATION

SLIS Student Research Journal

Volume 4 | Issue 1 Article 4

May 2014

Digitization and Digital Preservation: A Review of the Literature

Stephanie Routhier Perry
San Jose State University, srouthierperry@gmail.com

Digitization is rapidly becoming one of the standard forms of preservation for libraries, archives and information centers' analog materials. This newer process is allowing preservationists to ensure information contained within fragile, organic materials will still be viewable to future generations. However, as technology changes, there are concerns that the methods used today to preserve these materials are not going to be sufficient or even viable in the future. Software and formats change very quickly, and could be obsolete in a relatively short time period. This applies both to hard copy materials that are converted into digital copies, as well as born-digital items, or those who were created as digital copy initially. For this reason, digitization is not strictly a preservation activity, as the new files will require preservation as well. It is important to understand what digital preservation is, and how it can be effectively used to preserve collective knowledge for future generations.

There is a wealth of information on this topic in the literature today, and finding relevant articles and sources is not difficult. There are different schools of thought on digital preservation; some see it as the most important advancement in the topic of preservation, while others feel that it is not the only or even the best solution to keeping information safe. While there are opposing views on some aspects of digitization and digital preservation, there are also areas where most authors seem to agree. Budgetary issues, professional education, and increased technological currency are frequently mentioned as challenges in the field, and the need for better and more in-depth education, cost-sharing initiatives, and cooperation are universal suggestions.

This is a field where changes happen very fast. What was current at one time becomes out-dated quickly. The objective of this paper is to define the differences between digitization and digital preservation, and to get a broad overview of the current state of digital preservation. There are many subtopics within this field, and each of them are worthy of more in-depth study. However, this review will discuss the topic as a whole.

What are Digitization and Digital Preservation?

Digitization is the conversion of traditional, analog materials such as books, maps, and other paper items into an electronic, digital copy. This is not to be confused with digital preservation; as Conway (2010) notes, "It is important to establish clear distinctions between the terms 'digitization for preservation' and 'digital preservation'" (p. 64). Digital preservation is the conservation of all digital materials, whether they were born digital, such as emails, websites, videogames, and other electronic files, or whether they have been digitized from analog materials (Conway, 2010). Although digitization is often seen as preservation, this is not the case. According to Smith (2007), "Much is gained by digitizing, but

permanence and authenticity…are not among those gains". Digitizing may allow greater access to an artifact, but it comes with its own challenges.

Digital records have many the same function requirements that paper ones do. In other words, the work required by professionals to transfer, process, store, and preserve files, which allows a user to retrieve the desired information, applies both to analog and digital records (Galloway, 2009). However, of the main difficulties surrounding digital preservation compared with that of paper-based records is that the analog materials are much easier to maintain. Apart from storage and maintenance concerns, if a library or archive can keep its materials in an appropriate facility, files are relatively easy to preserve. Once an item has been digitized, however, that new version requires continuous, ongoing maintenance for as long as the record is to be kept. This presents huge cost and time implications for the facility (Sanett, 2013).Additionally, when it comes to digital materials, there is a large difference between storage and preservation. Storage is simple, as there is enough space in hard drives or in the cloud for as much material as can be created. The difficulty is that even if the stored data is intact, it may not be available or accessible, due to technological changes or human error in naming conventions. Preservation, that is, keeping the information available and usable for future generations, requires much more complex actions (Brand, 1999).

Differences in digital preservation needs occur depending on the type of institution involved. Libraries usually have published materials, which mean there is often more than one, and usually many, copies of the same artifact. Preserving the original material is important, but the information contained within is often more important to users (Galloway, 2009). Conversely, archives contain material that is unpublished and one of a kind. In this instance, both the information contained and the original artifacts themselves have intrinsic value. It is important for both libraries and archives to preserve the authenticity of an artifact, as well as the source of the object, the questions surrounding where it came from, whom it came from, and in what context it was created (Galloway, 2009).

Digitization and therefore digital preservation are no longer emerging tools; they are now the preferred and accepted practice for saving many analog records. Libraries and archives are in a transitional period and many are moving away from print into a primarily, or, in some cases, entirely, digital format (Moghaddam, 2010). Users expect instant access to materials wherever they are, and the only way this can be achieved is by digitization (Conway, 2010). It is important that future users will be able to access the information, be able to tell whether the information is accurate and preserved as it was intended to be, and use it in their intended way (Gladney, 2009).

Challenges

Although the act of digitizing analog materials, and the ensuing preservation of those materials, has many benefits and much to offer, it is not without challenges. Some of the main concerns include issues involving human error, data loss, fading memory, lack of effective education, and technological obsolescence (Kastellec, 2012). Despite the ready availability of mobile technology and hardware devices, digital resources are both human and machine dependant, which is perhaps one of their greatest limitations (Moghaddam, 2010). Knowing what to preserve, and the best method to use, is a major concern for professionals, and one that requires specialized training.

Changing Formats and Obsolescence of Technology

In addition to the technical difficulties, many of the issues surrounding digital preservation involve an institution's willingness to manage their digitized data (Teper, 2005). It must be acknowledged that that digital technology is not the only answer, and that hard copies of artifacts and materials are still very important. Acid-free paper and microfilm can last for 500 years in a library that is air-conditioned and dry, but digital files become obsolete every few years at best (Brand, 1999).

Quickly changing digital technologies are one of the most serious challenges associated with digital preservation. Because there is no way of knowing what formats or procedures people will be using 50 or 100 years from now, it is more important to focus on a shorter time frame. If professionals focus on the coming five, ten or even twenty years, they are more likely to have a better idea of what the practices will be (Gladney 2009). Software corporations are both part of the problem and the solution, as due to commercial interest, the devices, software and formats used to store information are often designed for obsolescence. In other words, corporations that design these methods design them with a shelf life in order to ensure the sales of latest upgrades and models (Reyes, 2013)**.** Digitization has many useful benefits, but if professionals cannot devise useful storage and retrieval techniques, there is the fear that future historians may find the current period a "dark age" of information from which little has survived.

One of the most significant events in an artifact's history is when there is a transfer from the original format to a successor. Any efforts to reformat an object and turn it into a digital file will result in the loss of some of the value of the original; therefore, care must be taken to ensure that the correct procedures are followed. All changes to files, once created, must be documented and embedded into the metadata in order to preserve the integrity of the information (Gracy & Kahn, 2012). Additionally, each successive record must have a copy saved for

archival purposes before the record is used, in order to ensure each step in the artifact's life has been preserved. Once a file has been changed, a signature or Digital Resource Identifier (DRI) should be used, especially for items that have more than one version being preserved over time; for example, updated versions of the same file (Gladney, 2009). Because these considerations are subjective, a framework is necessary, and professional associations are developing suggestions on this matter, including the code of ethics developed by the American Institute for Conservation (AIC). This code outlines actions that a preservationist should follow, such as to "strive to select methods and materials that, to the best of current knowledge, do not adversely affect cultural property or its future examination, scientific investigation, treatment, or function" (AIC, 2013, para. VI).

The US National Archives & Records Administration (NARA) holds that if something is worth preserving digitally, it must be preserved as close as possible to its original state (Galloway, 2009). This will guarantee the authenticity of the item and preserve it for historical study. It is also best to work with both digital and hard copy mediums to ensure safekeeping of all types of knowledge contained in a material, as digitization is only possible if works have been preserved in a more traditional format as well. Preservation strategies are more important that the actual formats used, and migration strategies for formats must be decided upon (Tennant, 1999).

There are a number of ways professionals can ensure that digitally preserved materials remain usable. It is important not only to preserve the record itself, but also the hardware and software it was created on and designed to be used with. Digital materials are very complex, and compression, encryption, and HTTP links that were active in original works make it much more difficult to extract meaning from a preserved artifact. Because it is often not possible to preserve a digital file exactly as it was when it was created or when it was analog, at the very least the focus must be on preserving the 'essence' of a file (Zorich, 2007). Emulators are computer programs designed to mimic or "emulate" other operating systems, and are one solution to outdated software or hardware (Galloway, 2009). Using an emulator allows users in the future to see exactly how the material would have looked, and by creating a similar operating environment to the original, helps the files to remain interactive. Because it is important to be able to see something in its original form whenever possible, in order to preserve historical authenticity, emulators would need to be created and updated regularly, as older technology becomes obsolete (Moghaddam, 2010).

When working with digital materials, there is the problem of mutability. An example of this is with a videogame. Watching the game is a very different experience than playing it, and it is difficult to get the full feel of the material when not using it in its intended context. Because of this, it is very important to

use the most descriptive metadata possible, in order to give future viewers a better understanding of what they are seeing (Brand, 1999). The format, the naming convention, and the reasons why certain aspects have been chosen will all likely be of interest to future historians, just as aspects of old artifacts, such as the binding used in an old book, or what was used for ink, are important to scholars today (Berger, 2009).

Untrained Staff and Human Error

Currently, humans are still performing most of the necessary acts of digital preservation. With human actions come human error, however, and many library and archives staff members do not have the required training and skills needed to carry out this important work. The professional skills needed include technical proficiency in areas such as encryption, metadata schema coding, and authentication, as well as traditional archiving skills which include cataloguing and classification (Sanett, 2013).

Despite an increase in library and information science (LIS) students taking digitization and digital preservation courses, many institutions do not have the budgetary funds for an archivist or trained librarian on staff, leading to increases in human error (Maberry, 2013). Using staff that have not been formally educated in digital archiving comes with risks. Photographic and audiovisual collections are especially prone to deterioration, and require specialists and an often large budget to care for them properly (Gracy & Kahn, 2012).

Digital preservation often competes with the traditional preservation of the original artifact, which leads to difficult decisions. With an increased demand for instant access to electronic records, administrators have to decide where to spend more money – on preserving the original material, or on ensuring instant access to a digital surrogate. Both types of preservation have similar requirements concerning time, expertise and budget, and is often a difficult choice for staff members when deciding where to focus their attentions (Gracy & Kahn, 2012). This issue requires properly trained staff, because a large portion of an organization's budget already goes toward the cost of facilities management, and considerations for off-site storage must be considered. Understanding what the most important artifacts are, and which should have precedence when preserving, is a difficult subject and will be different according to the needs of each institution and its users. With such an exponentially large and always growing bank of digital information, knowing what to save first is a difficult task (Maberry, 2009). It is important that professionals understand that when they make a choice on what to preserve first, they are also inadvertently choosing what might not be saved at all (Teper, 2005). Preservationists must determine what value an item or artifact currently has, and what it is likely to be in the future; taking content,

historical value, artifact condition, and the rarity of the item into consideration are important steps in choosing artifacts (Berger, 2009). Often, the intrinsic value of a book with all of its original information is preserved only when something has been deemed a "classic". However, this is a difficult task, and one that is subjective. How can one know what will be considered a classic in the future? Who decides whether something is important enough to preserve?

Additional parts of a book or other physical item, such as the jacket, typeface, paper, binding, and ink are all important parts of the artifact. When an artifact is digitally preserved, this information is lost, along with some of the intrinsic value. Reformatting also involves making choices about the text or artifact itself. The preservationist must decide what to put into a digital copy and what to leave out. Regardless of the choices made, this process will inherently change the document's meaning (Bee, 2008). Reformatting, or migration, is often necessary as a last resort, as hard copies will not last forever; however, it is important to note that any time a text is altered; it becomes more prone to error. As new technologies emerge and items are migrated repeatedly, it is not hard to imagine a product that hardly resembles the original (Bee, 2008).

Staff training and education are essential when digitizing or digitally preserving materials. A major problem with LIS educational programs has traditionally been a lack of cohesiveness; that is, different schools have taught different methods, and there has been no generally recognized framework (Sanett, 2013), and (Dubin, et al., 2009). This is changing, however, and the Society of American Archivists (SAA) has created a list of seven core competencies that a digital archivist should have, which includes the ability to communicate the requirements related to digital archives, to formulate the strategies needed to best organize and preserve them, and to "Integrate technologies, tools, software, and media within existing functions for appraising, capturing, preserving, and providing access to digital collections" (Society of American Archivists, 2013, para. IV).

Authenticity and Reliability of Material

As noted, when artifacts undergo a digital preservation process, there is the risk that the preserved item will eventually not even resemble the original. It has been noted that when proper procedures are carried out, digitization can be 99.65% accurate (Bee, 2008). While this figure is high, it still leaves room for error. Over time, as records are continually reprocessed, inaccuracies will compound on top of each other. Additionally, external and internal attacks on items, such as those by hackers, upset employees, or others committing acts of fraud or revenge, as well as economic failures such as a lack of funding or mission change, can all have a negative impact on materials (Gladney, 2009). Digital files are much more

difficult to classify as authentic, as it is very simple to manipulate a digital file and any changes are not always easily apparent (Smith, 2007). There is a similar issue with large-scale digitization projects, particularly when moving items from one repository to another. Information can be easily corrupted or lost (Dubin, Futrelle, Plutchak & Eke, 2009). During the course of a large project, many errors or losses can occur, but it is often too expensive and time consuming to check every part of every record. Staff members need to ensure they are taking all necessary measures to preserve the integrity of the records they are working with.

Standardization

Another major issue with digital preservation education is that the metadata used in digitization is often not standardized, and different variations of the same word or description can cause errors. This is not as critical when it is humans doing the preservation, but as processes become more automated, complications will arise. Because algorithms and software programs can only rely on the information inputted into them by humans, metadata must be standardized so that nothing is lost or left to chance (Dubin, et al., 2009). Although there is no universally agreed-upon set of standards, some are becoming more standard in the industry, particularly in North America. Two of the most prominent are Preservation Metadata: Implementation Strategies (PREMIS) and Metadata Encoding Transmission Standard (METS) (Sanett, 2013). Each has recommendations on metadata standards and guidelines for use.

Although metadata schema, best practices, and professional education have been varied, this is changing, and the Library of Congress's Digital Preservation Outreach and Education Program, which aims to build collaboration between LIS schools and professionals, is one example of the way in which the profession is recognizing that training must be more widely available and more cohesive (Sanett, 2013).

There are many types of standardized repository frameworks around the world, such as the North American Open Archival Information System (OAIS), and the European Repository Toolkit (Galloway, 2009). Repositories using the OAIS framework should ensure files are preserved with needs of both the preservationist and the end user in mind. There are a number of rules and suggestions for archives following this format, including the need to determine the scope of the archive's user community, ensure that the preserved information is independently understandable to the users, and make the preserved information available to the users (Galloway, 2009). The Repository Toolkit allows institutions to integrate digital preservation tools into already existing repository software.

Copyright and Cost Issues

In keeping with other library and information issues, copyright is always a concern. There are many uncertainties in the copyright laws both nationally and internationally, and it is a very grey area. There are complicated questions when examining the copyright of digitized items. Is the material that has been digitized considered published or unpublished? Does the original creator own the rights, or the preservationist? These questions are difficult to answer (Kastellec, 2012). Time can be a factor with copyright issues, especially for special collections or rare items. Trying to find a rights-holder can be time consuming and expensive, and is often an unsuccessful exercise, particularly when working with older materials (Gracy & Kahn, 2012). One implication of personal records being digitized or created as born digital files is that they are then stored in repositories or cloud storage, and therefore beyond the control of the creator. Ingestors of material can ask a server to remove personal records, but there is no guarantee they will do so. Conversely, a record could be deleted at any time; again, without the consent of the ingestor. Ensuring access to records while maintaining privacy is a problem that requires further study (Reyes, 2013).

Finally, as with many issues surrounding libraries and archives, cost is a key issue. Libraries, archives and storage repositories generally do not pay for themselves or have any immediate financial benefit to a parent institution, so funding is always a topic of concern (Brand, 1999). Because digital preservation is a relatively new area, it is difficult to gauge what the cost to departments and institutions will be when using this method of preservation. There are a large number of costs to factor in, such as the cost of program and project management, skills training for staff, and the new software needed to implement the retention of digital assets (Sanett, 2013). The amount of information available that could be preserved is much more than professionals can realistically save, and cost is always a limitation. This financial reality underwrites all other factors when it comes to challenges. No matter what professionals may want to do, or what the user needs are, money is at the root of all decisions (Kastellec, 2012).

The problems associated with digital preservation are often too great for one organization alone to handle, and the cost can be prohibitive. Fortunately, libraries and archives do not have to shoulder this burden alone; organizations can collaborate in order to manage their collections together (Zorich, 2007). Sharing off-site storage via a consortium is one way to lower the cost of keeping hard copies. In this way, if a library or archive cannot afford to store important artifacts correctly, they can send them to an archive or central repository that can. By having the original materials in offsite storage, each member of the consortium can access the digital copy and share resources (Gracy & Kahn, 2012).

An increasingly common repository framework is the Lots of Copies Keep Stuff Safe (LOCKSS) system, which allows a group of consortia members to have a group of interconnected LOCKSS boxes, which continually monitor each file in each box, ensuring nothing has changed with any of them (Gracy & Kahn, 2012). Because authenticity and reliability are always important considerations when archiving and preserving any material, at least one copy of every file or piece of information should be stored in a safe location where it is never accessed or attached to any other computer network. This is known as a dark archive, and should be a collection of materials that do not need to be ever be changed or manipulated (Gladney, 2009). There should be more than one of these "master copies" to ensure validity if needed. Security is an important consideration for these repositories, both to guard the collections against malicious damage, loss, forgery, and theft, and to ensure files are presented according to users' needs (Gracy & Kahn, 2012).

As always, cost is a consideration when developing a shared repository framework. Organizations must decide in advance what the outline for cost distribution will be, and how the program will be managed. Digital storage itself is inexpensive, but managing it properly is not (Stewart, 2012). In addition, it should be remembered that although consortiums are a good way to handle the costs and challenges present, it is important for institutions to develop their own digital preservation policies, in order to be able to move forward in an organized, manner that suits all parties (Sanett, 2013). Consulting with similar institutions or those in the area or with the same or similar mandate would be beneficial.

Awareness

It is imperative that professionals follow best practices in all aspects of digital preservation, including maintaining and preserving the authenticity of the digital objects, making backups, generating appropriate (and standardized) metadata, and continuing to track all surrogates.

Naturally, the above considerations are only viable if professional staff and funders are aware of the problem. It is vital that librarians and archivists make the public aware of the importance of digital preservation. In recent years, government and private bodies have begun to recognize the importance of archiving digital information, and some initiatives are already underway, such as the creation of the National Digital Information Infrastructure and Preservation Program (NDIIPP). This $99 million program, created by the United States federal government in recognition of the importance and prevalence of digital archiving, has the responsibility of creating a national program that focuses on digital preservation of information (Teper, 2005).

Because there is such a large wealth of information that is now digitized, it is almost impossible for libraries and archives to save it all. Up to 80% of websites are updated and/or gone within one year of creation; this makes it nearly impossible for professionals to keep up (Lasfargues, Martin & Medjkoune, 2012). Instead of trying to save everything, some believe that the focus should be on devising the correct ways of preservation and the tools that should be used, and then teaching them to others. Particularly in regards to personal information and records, the public needs to understand the implications of information loss, and take steps to preserve their own information, thereby contributing to the collective historical record (Reyes, 2013). Corporations, governments, and even private citizens should be in charge of their own records, with input from library and archival professionals (Galloway, 2009).

Conclusion

There are different schools of thought on digitization; some see it as the most important advancement in the topic of preservation, and others feel that while it has a place, it is not the even the best solution to keeping information safe. Despite these different attitudes, digitalization, and the subsequent preservation of the digitized files, is an important topic in the world of libraries, archives, and other information centers. As this is a dynamic topic that changes often according to new information and techniques, it is one that will need to be constantly re-addressed. Librarians and archivists have two jobs when it comes to digital preservation; they have to keep up with existing practices, while still finding ways of making digitization happen. As well as this, it is important to focus on the future, anticipating what challenges could arise, and devising solutions. As with all topics surrounding the LIS field, it is important to remember not just the information that needs to be preserved, but also the needs of the information seekers. Librarians, archivists, and other information professionals must continually strive to ensure that users will be able to access the preserved information.

References

American Institute for Conservation. (2013). *Code of ethics*. Retrieved from: http://www.conservation-us.org/about-us/core-documents/code-of-ethics#.U3UeWdJdV8E

Bee, R. (2008). The importance of preserving paper-based artifacts in a digital age. *The Library Quarterly, 78*(2), 179-194.

Berger, S. (2009). The evolving ethics of preservation: Redefining practices and responsibilities in the 21st century. *The Serials Librarian, 57*(1), 57-68.

Brand, S. (1999). Escaping the digital dark age. *Library Journal, 124*(2), 46-48.

Conway, P. (2010). Preservation in the age of google: Digitization, digital preservation, and dilemmas. *Library Quarterly, 80*(1), 61-79. Retrieved from http://www.editlib.org/p/106837

Dubin, D., Futrelle, J., Plutchak, J., & Eke, J. (2009). Preserving meaning, not just objects: Semantics and digital preservation. *Library Trends, 57*(3), 595-610.

Galloway, P. (2009). Digital archiving. In *Encyclopedia of Library and Information Sciences.* (pp. 1518-1527). Taylor & Francis. doi:10.1081/E-ELIS3-120044332

Gladney, H. M. (2009). Long-term preservation of digital records: Trustworthy digital objects. *The American Archivist, 72*(2), 401-435.

Gracy, K. F., & Kahn, M. B. (2012). Preservation in the digital age. *Library Resources & Technical Services, 56*(1), 25-43.

Kastellec, M. (2012). Practical limits to the scope of digital preservation. *Information Technology & Libraries, 31*(2), 63-71.

Lasfargues, F., Martin, C., & Medjkoune, L. (2012). Archiving before losing valuable data? Development of web archiving in Europe. *Bibliothek Forschung Und Praxis, 36*(1), 117-124. doi:10.1515/bfp-2012-0014

Maberry, S. E. (2009). Archiving 2.0: Problems, possibilities, and the expanding role of librarians. *Art Documentation, 28*(1), 40-43.

Moghaddam, G. G. (2010). Preserving digital resources: Issues and concerns from a view of librarians. *Collection Building, 29*(2), 65-69. doi:10.1108/01604951011040152

Reyes, V. (2013). We created it, now how do we save it? Issues in preserving personal information, a review. *Preservation, Digital Technology & Culture, 42*(3), 150-154. doi:10.1515/pdtc-2013-0020

Sanett, S. (2013). Archival digital preservation programs: Staffing, costs, and policy. *Preservation, Digital Technology & Culture, 42*(3), 137-149. doi:10.1515/pdtc-2013-0019

Smith, A. (2007). Digitization is not preservation – at least not yet. In *The whole digital library handbook* (Preservation). Retrieved from http://site.ebrary.com.libaccess.sjlibrary.org/lib/sjsu/docDetail.action?docID=10194656

Society of American Archivists. (2013). *DAS Curriculum Structure.* Retrieved from Digital Arechivists Specialist (DAS) Curriculum and Certificate Program: http://www2.archivists.org/prof-education/das/curriculum-structure

Spindler, R. P. (2009). Electronic records preservation. (pp. 1682-1688) Taylor & Francis. doi:10.1081/E-ELIS3-120008662

Stewart, C. (2012). Preservation and access in an age of E-science and electronic records: Sharing the problem and discovering common solutions. *Journal of Library Administration, 52*(3), 265-278. doi:10.1080/01930826.2012.684505

Tennant, R. (1999). Time is not on our side: The challenge of preserving digital materials. *Library Journal, 124*(5), 30-1.

Teper, T. (2005). Current and emerging challenges for the future of library and archival preservation. *Library Resources & Technical Services, 49*(1), 32-39.

Zorich, D. M. (2007). Preservation and access for a digital future: The WebWise conference on stewardship in the digital age. *Curator, 50*(4), 455-460.