More Product, Less Procedure and Digitization of Archival Materials

Sarah Sammis

November 22, 2011

Michael Hooks

LIBR 256-11 – Archives & Manuscripts

San José State University

Abstract

Greene and Messiner's more product, less process (MPLP) approach to tackling archival backlog can also be applied to the digitization process of archival materials. As more people have access to computers and the internet, there is growing demand for online sources of quality, trustworthy research information. Archives and special collections can and should leverage their materials to fill the growing research demand.

Digital records, though, call into question some of the tenants of archival work, namely, provenance, authenticity and original order. The digitization process can also potentially damage delicate items through tearing and warping. Digitization also takes time, a precious commodity that many archivists already feel is limited. There are, though, steps to take to avoid damage, protect the tenants of archival work. These steps are well within the bounds of MPLP.

Keywords: digitization, MPLP, digital surrogacy

Introduction

Panofsky (1974) in the introduction to *Meaning in the Visual Arts* writes, "[T]he records have to be 'decoded' and interpreted, as must the 'messages from nature' received by the observer. Finally the results have to be classified and coordinated into a coherent system that 'makes sense'" (p. 7). While he's talking about "humanists" and not archivists, per se, he has described the archival process beautifully. In this modern day of computers, sharable metadata, and searchable databases, I argue that the classification Panofsky mentions is the standard processing (assessment and description) already in place while 'coordination' is the digitization of analog materials. Panofsky continues: "To grasp reality we have to detach ourselves from the backlog" (p. 24). In the context MPLP, archivists must detach themselves from the reality of the backlog to tackle the processing which *should* now include some level of digitization.

To study the relationship between digitization and the processing of archival materials, one must understand the terminology. For the scope of this paper, much of the complexity of digital archiving has been removed. This paper will not cover extensively "born digital" materials nor will it look in any depth at the problems of copyright protection, privacy issues or specific approaches to description of materials with metadata.

Digitization is "distinguished from 'data entry', which is the process of typing textual records" (Pearce-Moses, 2005) and while the information may start out in an electronic format (such as magnetic tape), the digitization process involves turning information into a format that can be stored, searched and retrieved via a computer or web enabled device.

Born digital materials did not have to go through the digitization process to get into their current format (Pearce-Moses, 2005). These can be emails, digital photos, word processing documents, web pages and any other computer data.

Digital surrogates are digital copies of analog archive documents. They are not meant to replace the original, but serve as a tool for researchers working remotely. Surrogacy can also help in the preservation of fragile or damaged documents. These digital surrogates will also need preserving within the context of the archive or special collection (Capell, 2010).

Digital archiving is the care of digital information, either which was born digital or has been digitized. Digital archiving is still a relatively new form of archiving. It is also a hotly contested field, with those arguing for the digitization of everything and others asking for a moratorium until the technical and ethical issues can be sorted out and standards written. Realistically, digitization has happened, is happening, and will continue to happen and the best practices for it will arrive from the successes and failures of those who partake in the process.

Methods of digitization

The digitization process requires different methods for different types documents. Archives can have any number of things in their collections from papers, photographs, negatives, film, maps, paintings, music (on cylinders, 78s, vinyl, magnetic tape, CD) to three dimensional objects such as furniture. Each type of document requires a specialized approach and each document itself may require special care. These methods include: OCR (optical character recognition), photograph or negative scanning, paper scanning, digital photography of 3D objects, conversion of music to MP3 and conversion of film to mpeg or other digital video format.

Optical character recognition begins in the same fashion as photographic scanning, the scanning of the page containing the text. The scanned pages are then processed by a computer program that can recognize printed text and convert it into computer readable text. This conversion allows for full indexing of texts for online database searches. Maxwell (2010) cites Google Books as the largest example of OCR and praises it thusly: "No special software is required to use Google Books (n.d.) currently the most important resource for digital historical research."

The process though is problematic, not all letters are correctly recognized. Stray marks or dust on the original text might be interpreted as being part of the text. Good OCR requires human intervention. Another online archive of books that uses OCR is Project Guttenberg but it has a team of volunteers to proofread the OCRed text before it is published on their website. The downside of voluntary proofreading is that turnaround from original scan to completed book can take years, as I discovered when I managed one of these Project Gutenberg book conversions.

Image scanning, either scanning of photographs or negatives or original artwork has its own series of advantages and disadvantages. The primary benefit is preservation through digital surrogacy. Scanned images though don't lend themselves to automated indexing or metadata creation that comes with OCRed texts. The Robert Walker Photograph Collection used scanning to recover digital surrogates of negatives that were degrading beyond the point where they could be repaired or the damage stopped (Cappell, 2010). "The results of the test scans exceeded expectations" (Capell, 2010, p. 246) and allowed the preservation of 72 images that would have otherwise been lost.

As Angeli & Todors point out, good description of scanned images is a must. As these descriptions cannot be automated; they must be created by knowledgeable people. Finding experts who are also archivists can slow down the description process (Angeli & Todors, 2010, p. 11).

Three dimensional objects, music, and film and video will also have special handling requirements for digitization. For the purpose of this paper, I focused mainly on two dimensional items: things such as paper documents, printed photographs, photographic negatives and oversized flat items (maps, for instance).

Areas of Concern

As Cook (1997) notes, the goal of archivists is to "preserve the memory of the world" (p. 2). He continues to say that archivists exist to "make other people's work possible" (p. 7). The more product, less procedure process (Greene & Meissner, 2005) method and digitization should be in alignment with Cook's goals. Digitization, some argue, is against the other tenants of the archive: provenance, original order and authenticity. Maxwell (2010) worries that digitization will replace the desire of archives to preserve paper records. For him, the primary function of the library is "preservation, preservation, preservation" (p. 26). The remainder of the literature read for this paper, though, suggests that libraries and archives have a more complex function than just *preservation* that includes providing access to information and artifacts of historical and cultural value. That said, Maxwell's call for digital surrogacy instead of the "sacrifice of irreplaceable materials on the altar of digitization" (p. 26) is sound advice and is in keeping with the direction digitization appears to be taking within the archival setting.

Provenance is the "original source of something" as well as the information regarding those origins (Pearce-Moses, 2005). As provenance provides meaning to a collection, Monks-Lesson (2011) asks if the flexibility of internet archives works "against the archival principle of provenance" or if it can be redefined to fit the "flexible context" of digitized materials (p. 38).

To preserve provenance, the materials of one collection should be kept separate from the materials of other collections. This separation helps to "preserve their context" (Monks-Lesson, 2011; Cook, 1997). The no-mixing taboo though contradicts one of the key uses of a digital archive: search results that can be drawn from multiple online collections. The UCLA Film and Television Archive (Archive Research and Study Center, n.d.) allows one to preserve provenance by searching specific collections, if the collection name is known.

Searching of digitized archive materials has been a feature of digital archives since the beginning. Herschler & Slany (1982) describe in their article how search was built into the database used by the Foreign Affairs Processing Center. While the specific methods used to automate digitization and to search generated records have changed (telegrams and primitive email) the basic workflow is still recognizable. Like the UCLA Film and Television Archive, the FAIS searches could be set to specific collections or pieces of collections depending on security clearance.

The next tenant of archival work is authenticity. Authenticity is the "quality of being genuine, not counterfeit and free from tampering" (Pearce-Moses, 2005). Put simply, archivists don't want to waste time and space storing fakes. Digital records, whether born digital or digital surrogates, can be altered. Purists worry that the authenticity of a digital record can never be fully trusted (Casewell, 2009; MacNeil, 2011; MacNeil & Mak, 2007, Monks-Lesson, 2011). Having the originals on hand to check for alterations is a must (whenever possible) to preserve the trust in the collection.

Original order serves two main purposes: it preserves "existing relationships and evidential significance" while saving the archivist the time needed to create a new ordering in the processing (Pearce-Moses, 2005). Boles (1982) notes an uneasiness among archivists of personal papers and government records for anything that can disrupt the original order of a collection, even if that re-ordering is done on the fly on a page of search results. Holmes (1964) while not speaking specifically of digitization, argues that the item level processing ("flattening and microfilming") can (and perhaps *should*) be "done by subprofessional employees, or by lower grade — that is, less experienced — professional workers" (p. 38). Assuming that the materials being processed aren't in poor condition, aren't rare or in poor condition, basic scanning or digital photography (of three dimensional objects) could in fact be done by archive interns, volunteers or entry level staff, thus saving time for the archivist while still producing searchable, digital information that researchers can use.

Besides the search results page, there is another way digital materials can be re-ordered, and that is through the sharing of data via XML (whether presented as an RSS feed or an EAD search tool). Monks-Lesson (2011) notes that records take on new meanings and contexts depending on how they are ordered. While the original order and context "should be preserved", new contexts should be allowed and the "'fluidity, flexibility and ultimately uncontrollable nature'" (p. 57) of shared information should be explored to foster research.

Digitizing of materials, though, creates new concerns over copyright. Part of digitization is the copying of an object to move it from one medium to a digital medium. Digital media by its very nature can be copied. In the meantime, copyright protection has been expanded, keeping some materials out of the public domain for decades longer. There is enough literature on copyright to warrant a separate paper. For the purpose here, I include copyright merely as a concern that is causing some archives to hold back on digitization (Greene & Meissner, 2005; Prom, 2011; Riley & Shepherd, 2009).

More process, less procedure and some digitization

For those archives that do chose to digitize, there are practical concerns about the process: mainly in terms of time and the potential damage that can be done in the scanning process. The Library of Congress (1999) while generally for digitization of collections, suggests that archivists proceed with caution when beginning a new digitization project.

For each new project, the LOC recommends that archivists experienced with the collection "assess the fragility of and risks to the originals" to design digitization procedures that will "help protect the originals though the handling they will receive during scanning" (p. 3). Archivists should customize scanning projects to meet the specific needs of each collection, rather than trying to create an umbrella procedure. While it is tempting to race through a digitization project, it is better to pick techniques (even slower ones) that will better preserve the materials during digitization.

Damage to materials can include ripping and tearing and distortion due to heat buildup. Leather, parchment, film and photographic prints are especially vulnerable to heat. The Library of Congress recommends "adequate cooling and air-circulation to counteract heat that builds up from the equipment, and lights when using overhead capture devices" (LOC, 1999, p. 3).

Archivists trying to process a backlog, might be worried that the assessment of each collection for digitization will further add to the time needed, thus undoing any net benefits of MPLP. Samouelian (2009) and Greene and Meissner (2009) suggest finding a balance between user demand for digital surrogates and shareable metadata and the "realities of managing growing collections" (Samouelian, p. 46). In the MPLP paradigm, the level of scanning should strive not for "perfect" but for "good enough" (Greene, 2010). Archivists should match the level of their work with the collection's "relative value and importance" for both description and digitization. First hand knowledge of the collection through the accessing and processing will help in determining the level of digitization necessary (a few select items, all of the items or perhaps just a searchable description of the items) (Riley & Shepherd, 2009).

Just as the MPLP suggests a compromise between the time taken to process collections and the amount of processing accomplished, I suggest a compromise between no digitization and limited digitization that can leverage the advantages of digital surrogacy. Greene and Meissner (2005) outline the benefits of MPLP thusly: "1) expedites getting collection materials into the hands of users; 2) assures arrangement of materials *adequate* to user needs; 3) takes the *minimal* steps necessary to physically preserve collection materials; and 4) describes materials *sufficient* to promote use" (p. 212-3).

Digitization serves those four tenants of MPLP very well. It gets the materials into the hands of users through searchable databases and websites; it allows the user to rearrange (through search results, bookmarking or tagging) the materials to suit their research needs *without* affecting the actual order of the collection; if the description of the collection is produced on a computer it can be put online with *minimal* effort; and finally these online descriptions will promote use by making them easier to find, even by those who cannot physically come in to see use the collection in person.

For all the trouble of digitization: concerns about provenance, authenticity and original order; concerns about copyright; potential damage to materials and the extra time needed to digitize, is the result worth the effort? The literature review says yes. The benefits to digitization include: greater access to materials, full text search, shareable metadata, RSS and access to materials without the originals having to be touched and potentially damaged, lost or misfiled.

Access is the primary benefit of digitization. The "proliferation of online archive websites" make "archival material visible to anyone and accessible in virtual form to everyone" equipped with a web-enabled device an internet connection (Monks-Lesson, 2011). With online access comes online searching, a tool that helps researchers find and identify information relevant to their projects more easily (Fear, 2010).

Archives that provide sharable metadata either as RSS or EAD open their access even further by allowing collaboration between archives as well as the mixing of files. Marchioni and Maurer (1995), cited by Chen and Chen (2010) argue that the sharing of expensive resources benefits all libraries (and archives) by helping bring "people and ideas together" (p. 7). If libraries and archives are the repository of culture, the open sharing of data through digitization and metadata description can bring together these different pieces of the cultural repository to paint a broader and more complete picture (Monks-Lesson, 2011). As Dushay and Hillmann (2003) note, the sharing of digital surrogates can bring to light duplication but they argue this duplication can help archives improve their metadata and give researchers more resources from which to pick the best one for their needs.

Conclusion

Digitization is a process that archives can leverage to provide greater access to their collections while protecting rare, damaged or hard to handle documents. Digitalization allows for the searching and sharing of data between archives and in ways previously not possible. For the most part, digitization is used to create digital surrogates, or copies of original documents still preserved by the archive, but in some cases where the original is decaying, digitization can be used as preservation.

Digital records have brought forth concerns over provenance, authenticity and original order. While these problems still exist for born digital files within an archival context, for the purpose of digital surrogacy, they are addressed by the physical originals and how they are cataloged and described. Researchers and archivists should not conflate born digital files with digital copies. Where research requires complete confirmation of provenance or authenticity or exploration within the context of original order, the researcher should make arrangements to see the collection in person.

Digitization takes time to do properly, but not all collections need the same amount of digitization. The advice laid forth by Greene and Messiner for their more process, less procedure approach can and should be applied to digitization.

References

Angeli, M. M. and Todors, R. (2010). The Biblioteca Marucelliana and its database of prints and drawings: A work in progress. *Art Libaries Journal 35*(2). 8-12.

Archive Research and Study Center (n.d.). The UCLA Library: Film and Television Archive [online catalog]. Available at: http://cinema.library.ucla.edu/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First (accessed November 2011).

Boles, F. (1982). Disrespecting original order. *The American Archivist 45*(1). 26-32.

Capell, L. (2010). Digitization as a preservation method for damaged acetate negatives: A case study. *The American Archivist 73*(Spring/Summer). 235-249.

Casewell, M. (2009). Instant documentation: Cell-phone-generated records in the archives. *The American Archivist 72*(Spring/Summer). 133-145.

Chen, C.M and Chen, C.C. (2010). Problem-based learning supported by digital archives: Case study of Taiwan Libraries' history digital library. *The Electronic Library 28*(1). 5-28.

Cook, T. (1997). What is past is prologue: A history of archival ideas since 1898, and the future paradigm shift. *Archivaria 43*(Spring 97).

Greene, M. A. and Meissner, D. (2005). More product, less process: Revamping traditional archival processing. *The American Archivist 68*(Fall/Winter). 208-263.

Herschler, D. H. and Slany, W. Z. (1982) The "paperless office": A case study of the State Department's Foreign Affairs Information System. *American Archivist 45*(2). 142-15

Holmes, O. W. (1964). Archival arrangement — Five different operations at five different levels. *American Archivist 27*(1). 21-42.

LOC (1999). Conservation implications of digitization projects. Library of Congress NDLP and the Conservation Division. Available at: http://memory.loc.gov/ammem/techdocs/conservation.html (Accessed October 2011).

Maxwell, A. (2010). Digital archives and history research: Feedback from an end-user. *Library Review 59*(1). 24-39

Panofsky, E. (1974). *Meaning in the visual arts*. Woodstock, NY: The Overlook Press.

Pearce-Moses, R. (2005). *A Glossary of Archival and Records Terminology*. Available at: http://www.archivists.org/glossary/ (accessed November, 2011).

Prom, C. J. (2011). Using web analytics to improve online access to archival resources. *The American Archivist 74*(Spring/Summer). 158-184.

Riley, J. and Shepherd, K. (2009). A brave new world: Archivists and shareable descriptive metadata. *The American Archivist 72*(Spring/Summer). 91-112.

Samouelian, M. (2009). Embracing Web 2.0: Archives and the newest generation of web applications. *The American Archivist 72*(Spring/Summer). 42-71.