

Assignment 2:

Scanning, Digitization, and Imaging

David Gwynn

LIS 630 - Dr. Anthony Chow

27 April 2008

TABLE OF CONTENTS

INTRODUCTION	1
SCANNING, DIGITIZATION, AND IMAGING BASICS	1
1. Photographs and Image Scanning	1
2. Books and other text-based materials	2
3. Maps	3
4. Audiovisual materials	4
CURRENT AND FUTURE TRENDS AND ISSUES	5
1. Preservation vs. access	5
2. Prioritizing materials and selection criteria	6
3. Copyright issues	6
4. Ownership and licensing	7
5. Digital preservation and format compatibility over time	7
6. Other issues and concerns	7
CONCLUSION	8
REFERENCES	10

INTRODUCTION

Scanning and digitization of materials both for preservation and for access purposes is a major issue for libraries and archives at the present time. Photographs, books, maps, and audiovisual materials are the most likely candidates for digitization. Projects can be designed for preservation purposes, to facilitate access to materials, or both, with resulting differences in process and format. Other issues center around copyright, the suitability of digital materials as a long-term archival format, and prioritizing materials for digitization. This paper will attempt to provide a brief overview of the mechanics of digitization, and of some of the issues surrounding the concept.

SCANNING, DIGITIZATION, AND IMAGING BASICS

1. Photographs and Image Scanning

Photographs have been among the most common and visible targets of digitization projects to date, perhaps because photo projects can be started with minimal and relatively inexpensive equipment; a computer and a flatbed scanner are really all that are required to start a project involving flat photographic prints. High quality scans of negatives, slides, or film transparencies will require more equipment, either a film scanner or, at minimum, a transparency adapter that can be attached to a flatbed scanner. Sheet-feed scanners are also available and can be used for materials (documents in particular) that would not be damaged by them (CDP Scanning Working Group, 1999).

Conventional wisdom suggests that archival master copies be scanned at the highest feasible resolution and saved in a format that does not use compression algorithms; TIFF format at a minimum 600 dpi seems to be something of a consensus, although the CDP Scanning Working Group suggests a minimum spatial resolution of 3000-5000 pixels across the long side of the image rather than any specific dpi resolution (1999).

Once master files are completed, it is then possible to create lower-resolution, compressed files in JPEG format, which can then be used for online access. The much smaller JPEG files allow for much faster download times and compatibility with web browsers. Data about the files (including links between the JPEG "access" copy and the TIFF master, plus cataloging and other metadata) can then be stored in a separate XML data file (Koelling, 2004; Hughes, 2004).

With photographs (and other audiovisual materials, for that matter), the issue of image manipulation (color correction, etc.) is an important ethical concern. Preservation masters must involve as little of this editing as possible, although manipulating copies can be a valuable research tool as it can uncover material that may not be visible to the naked eye (Koelling, 2004).

2. Books and other text-based materials

Projects involving digitization of entire books date back to the Gutenberg project, established in 1971, and many such full-text digitization projects are now in progress. Examples cited by Schlumpf (2007) include the Google Books project and the Open Content Alliance (Microsoft, Adobe, and others). Other players include the JSTOR project (working with older academic journals) and numerous microfilm

digitization schemes both by major players such as Proquest and by local libraries and other institutions.

To minimize damage to bound materials, digital cameras rather than scanners are often used to create the original image scan of the page. There are varying degrees of automation to the process, including some use of mechanical page turners. As a general rule, the more fragile the materials, the less automation can be used. Digitizing books is also a slow process for nonprofits -- Google and Microsoft have better technology available to them - and is also a somewhat costly undertaking in general. According to the Internet Archive, about \$30 per title is average (Nickish, 2008). Kaplan (2007) mentions a figure of \$.10 per page.

Adobe's PDF, or portable document format, is becoming the most common format for digitized materials, as it allows both an accurate visual representation of the original and the potential for full-text searchability. Two steps are generally involved, the first being an image scan of the printed page, producing a visual copy, and then an OCR (optical character recognition) scan, which will capture the text. OCR scans can have a rather high error rate, and there has been some suggestion that simply retyping the text is sometimes more cost-effective than using OCR and subsequent cleanup.

3. Maps

Maps are another area of intense interest among digitization projects. One well-publicized example is the David Rumsey collection (<http://www.davidrumsey.com/>), a huge private collection being made available online by its owner. Map projects generally start with digital imaging as well (usually

through digital cameras rather than scanners), and the resulting files are then integrated with GIS (geographic information systems) software for searching, viewing, and manipulation (e.g. Google Maps “mashups”) online.

4. Audiovisual materials

As with photographs, the general idea for audiovisual materials is to create a digital master copy with as much information as possible, assuming that this master can then be used to create copies of a more manageable size for download and other access purposes. Both audio and video materials require considerably more equipment to digitize, and are also much more storage-intensive.

With audio, “resolution” is measured in terms of sample rate and bit depth, with a sample rate of 96 kHz and a bit depth of 24 being optimal, while 44.1 kHz and 24-bit are standard and acceptable. The minimal standard is 44.1 kHz and 16-bit, which is the quality of a commercial CD. The ideal storage formats for masters are uncompressed AIFF or WAV files (CDP Digital Audio Working Group, 2006). As with photos, there are numerous lower-quality, compressed file formats that can be created from these high-quality masters. Common compressed formats include MP3 and AAC/MP4, which are common online downloadable formats which offer sufficient audio quality for most applications.

Video is similarly defined by bit rate, frames per second, and compression. Relatively uncompressed formats (such as DV stream) are optimal for long-term masters, from which compressed copies can be made. Common compressed formats for online applications include the various MPEG formats (MPEG 2 is the standard for commercial DVDs) as well as Windows Media (WMV/ASF), Apple Quicktime (MOV), and

Flash Video (FLV); within these formats there are also additional codecs (compression and decompression algorithms).

CURRENT AND FUTURE TRENDS AND ISSUES

1. Preservation vs. access

Whether projects should focus on preserving originals or facilitating access is perhaps the biggest controversy surrounding digitization today. There is considerable discussion as to whether a digitized copy of a work is really a suitable means of archival preservation, since it is technology-dependent medium. Hughes (2004) states that digitization is “not a substitute for microfilming, and a digital master copy is not a ‘preservation master’” (p.51). There are also questions as to the suitability of digital surrogates from an ethical standpoint, since such materials can be easily manipulated (Koelling, 2004), and do not retain some of the “intrinsic” sensory qualities of the original (Westney, 2007).

“Access wins”, at least per a 2007 Online Computer Library Center (OCLC) report. As mentioned above, digitization to archival standards (assuming that it is appropriate at all) requires increased investment in both time and money. File sizes are larger and storage requirements are greater. So OCLC recommends focusing first on digitizing to increase access, transforming as much material as quickly as possible, and worrying about preservation, extensive cataloging, etc. later (Erway & Schaffner, 2007). In fact, digitizing for access can foster preservation, by providing a digital surrogate for use, relieving pressure on original materials (Schlumpf, 2007; Hughes, 2004).

2. Prioritizing materials and selection criteria

Selection and prioritization schemes for digitization projects generally focus on the value and uniqueness of materials, demand, the potential for added value through enhancement or increased access, intellectual property issues, technical feasibility, and preservation and safety of original materials (Columbia University Libraries, 2001).

Determining demand can be tricky; most agree that high demand items should be a priority, but there is also evidence that digitization itself can increase demand for items in a collection simply because it makes the collection more visible and accessible (Hughes, 2004). Manipulating demand in this manner can also be viewed as a revenue source (UNESCO, IFLA, and ICA, 2002) and a marketing tool to increase visibility and attract funding for other projects (Michel, 2005).

3. Copyright issues

Copyright is a significant issue, particularly with respect to whether or not digitized materials will be publicly available online. U.S. Copyright law generally permits library to make a “preservation” or backup copy of a copyrighted item for use within the library itself, and the general assumption is that a digitized copy would fall under these guidelines. However, providing access to the public outside the library (or at least outside a password-protected interface limited to library cardholders) is probably not permissible. Thus, digitization projects involving a publicly-accessible component must take copyright status into consideration. Of course, if the library holds copyright (say in the case of a donated photo collection) or if the material is

public domain, this is not a problem (Hoffman, 2005). Many of the titles being scanned by the Google Books and Microsoft projects are pre-1923 and therefore public domain (Kaplan, 2007).

4. Ownership and licensing

Ownership and licensing is distinct from the copyright issue in that it refers to who owns the digitized copies of the work, rather than the reproduction rights. This is not a problem with self-digitized projects, but it is controversial in cases such as the Google Books project, where Google retains ownership of the actual digital materials. For this reason, the Boston Public Library and others have opted for self-managed digitization rather than joining the Google project, relying significantly on volunteers in order to maintain ownership. Nearby Yale and Harvard, however, have opted for the commercial route, through Microsoft and Google, respectively (Kaplan, 2007; Nickish, 2008).

5. Digital preservation and format compatibility over time

An issue with digitization projects, and particularly with those designed for preservation or archival purposes, is the problem of format compatibility over time. Digitized materials must be maintained in a format that is usable with currently available software, and therefore should not rely on compression algorithms that may become obsolete. It is also necessary to consider a long-term migration strategy as storage media and formats change over time (Hughes, 2004; Feather, 2004).

6. Other issues and concerns

Other major issues surrounding digitization projects include ongoing funding issues, maintaining the project once the initial burst of enthusiasm wears off, OPAC

integration, and the question of use restrictions. Of these funding and budgeting is likely the most important and problematic; large-scale digitization initiatives are expensive and require tremendous commitments of time and resources, which is why OCLC and others recommend starting wherever it is possible to do so, and beginning to integrate digitization as an ongoing system rather than as a project (Erway & Schaffner, 2007). Schlumpf & Zschernitz (2007), Koelling (2004) and others also emphasize that it is possible to start small when faced with serious budget constraints.

CONCLUSION

It is clear that digitization is the wave of the future, particularly for archives and special collections. The mechanics are well-documented, and significant support resources are available. Fortunately, there is a high level of format standardization and agreement at this point across all media types. Controversies remain, however, in the areas of ethics, ownership (both copyright on original materials and licensing of digital ones), priorities and selection criteria, digital preservation, and the “preservation vs. access” battle.

OCLC recommends moving fast, urging quantity over quality where necessary, and focusing on digitization programs rather than on specific one-time projects (Erway & Schaffner, 2007). Adam Smith of Google echoes this all-inclusive approach and sense of urgency, stressing that “people are interested not just in searching web pages, but all the world’s information” (Nickish, 2008). The hint is that anyone

interested in getting started with digitization should perhaps look past the controversies and just get started.

REFERENCES

- CDP Digital Audio Working Group (2006). *Digital audio best practices, Version 2.1*. Retrieved April 23, 2008 from <http://www.bcr.org/cdp/best/digital-audio-bp.pdf>
- CDP Scanning Working Group (1999). *General guidelines for scanning*. Retrieved April 22, 2008 from http://www.cdpheritage.org/resource/scanning/documents/std_scanning.pdf
- Columbia University Libraries (2001). *Columbia University Libraries selection criteria for digital imaging*. Retrieved November 2, 2007 from <http://www.columbia.edu/cu/libraries/digital/criteria.html>
- Erway, R. & Schaffner, J. (2007). *Shifting gears: gearing up to get into the flow*. Dublin, OH: OCLC Programs and Research. Retrieved November 5, 2007 from <http://www.oclc.org/programs/publications/reports/2007-02.pdf>
- Hoffman, G. (2005). *Copyright in cyberspace 2: questions and answers for librarians*. New York: Neal-Schuman Publishers.
- Hughes, L. (2004). *Digitizing collections : strategic issues for the information manager*. London: Facet.
- Kaplan, T. (2007, November 9). Microsoft contracted to digitize library in 2008. *Yale daily news*. Retrieved April 25, 2007 from <http://www.yaledailynews.com/articles/view/22334>
- Koelling, J. (2004). *Digital imaging: a practical approach*. Walnut Creek, CA: Altamira Press.

- Michel, P. (2005). Digitizing special collections: to boldly go where we've been before. *Library Hi Tech* 23(3), 379-195.
- Nickish, C. (2008, April 22). Some libraries shun Google in book battle. *All things considered*. Audio clip retrieved April 23, 2008 from <http://www.npr.org/templates/story/story.php?storyId=89850150>
- Schlumpf, K., & Zschernitz, R. (2007). Weaving the past into the present by digitizing local history. *Computers in libraries* 27(10), 10-15.
- UNESCO, IFLA, and ICA. (2002). *Guidelines for digitization projects for collections and holdings in the public domain, particularly those held by libraries and archives*. Retrieved November 2, 2007 from <http://www.ifla.org/VII/s19/pubs/digit-guide.pdf>
- Westney, L. (2007). Intrinsic value and the permanent record: the preservation conundrum. *OCLC systems & services* 23(1), 5-12.