B Sides
http://ir.uiowa.edu/bsides

U of I SLIS Journal

05|14|2010

Learning From the Past: Digitization and Information Loss

Julia Skinner

Abstract:

The increased use of the Internet for research, as well as the desire to preserve information, has necessitated the digitization of library materials. This paper seeks to draw a comparison between the challenges of microfilming and digitizing, and what can be learned from previous formatting efforts to reduce data loss during current endeavors.

Keywords:
digitization | microfilm | special collections | archives | library science | libraries | information loss

Introduction

The increased use of the Internet for research, as well as the desire to preserve information, has necessitated the digitization of library materials. This paper seeks to draw a comparison between the challenges of microfilming and digitizing, and what can be learned from previous formatting efforts to reduce data loss during current endeavors. As reformatting remains the only long-term strategy for reducing data loss due to brittle paper, microfilm and digital reformatting must be considered an utmost priority (Chapman et al, 3). As this paper is only an overview of extant research, it does not seek to provide solutions or conclusions, but rather hopes to bring together a growing body of knowledge to serve professionals wishing to learn more about the current issues surrounding preservation in the digital age.

Reformatting to Microfilm: Challenges in Preserving Information

Microfilm has become the accepted format for the preservation of materials (Arms, 1). When microfilming efforts were initiated, several problems were encountered after information on microfilm came in danger of being lost. This was partly because of the creation of only one copy of the film, or the storage of film items in only one location. This meant if the available copy or copies were destroyed in a disaster, all record of that information would be lost. These same problems are being encountered in digitization efforts, with digital copies of documents only being held at one institution or on one machine, thus making them vulnerable to loss (Class notes, 2009). Quality checks also were not frequent with microfilm, meaning that information was lost due to poor-quality photographs (Baker, 45). In some cases, the microfilm is of materials that are set to be weeded from the collection, and so are sold or moved elsewhere (Baker, 44). An excellent history of microfilm and the challenges associated with the medium can be found in Baker (2000).

Importantly, despite some of the shortcomings associated with microfilming, digitization is still new enough that it has not made microfilm obsolete. As there is no universally agreed-upon technological approach or institutional capability that ensures continued access to digitized materials into the future, microfilm is still the preferred preservation format (Chapman et al, 2).

The main concerns with early digitization revolved around its use as a preservation format, as the longevity of digital documents was unclear (Chapman et al, 2). Although there are concerns with data loss, both microfilm and digital technologies are useful in part because they free up shelf space. However, digital technologies go one step further because they free up libraries from trying to digitize all their holdings to focusing their efforts on unique or valuable items within their collections (Class notes, 2009).

Information Loss in Digital Environments

Despite similarities in the problems encountered with digital and microfilm reformatting, digital objects both pose challenges and provide opportunities not offered by microfilm. One main concern, not present with microfilm, is the need for interoperability of different library systems that must work together to make materials accessible (IFLA, 1). If different libraries operate on incompatible platforms, they cannot share information, nor can users at one location access materials in another.

Issues also arise with the creation of metadata, a term referring to the tags applied to items which describe their content and attributes. No one metadata scheme has as of yet been widely accepted, although there are several (such as Dublin Core) which are available. The addition of metadata to a digitized image results in the ability to perform keyword searches, although some images scanned using OCR (Optical Character recognition) are fully text-searchable (Mindwrap, 2010). The main concern for metadata with digital items, as opposed to the more stable nature of MARC record searches, is that full-text searches produce an abundance of results, many of which may be irrelevant (as anyone who has ever used a search engine is well aware) (IFLA, 1).

As with most collection development and management policies, there are several challenges libraries face when deciding what should be made digital. The International Federation of Library Associations and Institutions (IFLA) offers the following guidelines for selecting items for digitization: collection strengths and unique collections should be the focal point of such a project, as these add otherwise inaccessible materials to digital holdings. Otherwise, institutions can work together to divide up the work of digitizing into manageable sections depending on staff skills and the available technical infrastructure, such as equipment, quality of internet connections, etc. (IFLA, 1). An *ad hoc* digitization program, wherein items are digitized and stored as they are requested, is not recommended as this produces a haphazard digital representation of the repository's physical holdings (IFLA, 1).

Avoiding the Loss of Information in Digitization Projects

Although reformatting documents into digital objects creates the possibility for information loss, there are things we can learn from mistakes made in the transition to microfilm, as well as some suggestions made by those experienced in the field. While microfilming does not inherently damage materials, items such as newspapers are cut from their bindings before filming (University of Georgia Libraries, 2008), which may increase the risk of damage if not done carefully. In order to prevent damage during digitization, it is important to use a scanner which does not harm the binding of a book, and these scanners have become available in recent years, although they are expensive (Arms, 1).

Another issue not adequately addressed in microfilming efforts is the creation of multiple copies. Just as backup copies are necessary in case microfilm is destroyed, they are equally important in the prevention of information loss in digital formats. It is recommended that at least three copies be made of each document: a store copy, a pull copy, and a use copy (Class notes, 2009). IFLA suggests that libraries may wish to consider working together to store redundant copies at designated institutions (IFLA, 1).

In addition to the creation of multiple copies, items should be stored in multiple formats. These could include Tagged Information File Format (TIFF), which is commonly used for the master copies of scanned items, Graphics Interchange Format (GIF,) which compresses files well without information loss, and Joint Photographic Experts Group (JPEG), a popular format, although compression sometimes results in loss. Several other options include Portable Document Format (PDF), which exactly displays recorded information, Plain Text (TXT), which produces small and easily portable files, and Hypertext Markup Language (HTML), which produces small files that can be formatted to display information in a desired layout on web browsers (McMillan Library, 1). It is important to select only those formats which have undergone extensive testing, are well documented, and can operate on a variety of platforms, which improves the sustainability of the information and reduces future maintenance costs (Beagrie, 2001).

Naming of digital items is important for retrieval of the correct document. The name must be unique, and must not be linked to its location. The importance of this tactic is illustrated by URLs, which specify the location of an item within the computer in order for it to be found. If the item is moved to a different part of the computer's memory, the ability to access this record is lost (IFLA, 1).

There are several possible ways to avoid information loss in this way. One is a PURL (Persistent URL), which is linked to a URL. This allows continued access to a document, even if the record is moved and the URL updated. Another is a Uniform Resource Name (URN), which provides a framework for naming digital objects through the use of authorities. Like a PURL, such records are linked to a URL, but unlike PURL, one URN can be linked to multiple URLs to reflect multiple locations or formats. Lastly is the Digital Object Identifier (DOI) System. This method was developed primarily for publishers to reliably identify and access materials to resolve intellectual property issues (IFLA, 1).

The obsolescence of digital formats is another serious problem facing libraries. One must be able to transfer information from an obsolete source (such as a floppy drive or my personal favorite, the laser disc) and store it in such a way that current equipment will be able to read the storage medium. Since storage mediums become obsolete quickly, libraries must continue to move data from medium to medium in order to keep up. This becomes an issue for information loss, and materials may become distorted or inaccessible. As of yet, there is no standard guiding data migration from one format to another (IFLA, 1). While not widely addressed, the cost of such reformatting and the purchase of new equipment might prove to be another barrier to libraries seeking to update their digital holdings.

There are also profession-wide changes that need to take place to minimize information loss. The first of these is increased inter-institutional collaboration, as discussed earlier, which would assist in the storage of redundant copies and increased information access. Standards also need to be implemented both for data migration and metadata, as already mentioned. Standards for archival quality for scanned images also need to be established (Chapman and Kenney, 1996), which would allow for greater quality control in the digitization process and avoid some of the problems with poor image quality that plagued the transition to microfilm.

As microfilm is still the standard for materials preservation, the switch to digital will not take place overnight. This is especially true as digital projects experience ‘growing pains’ in regards to compatibility with multiple platforms, a lack of standards, and the long-term integrity of data (as well as its continued accessibility). Because of this a ‘hybrid’ approach, where information would be both filmed and scanned, was recommended in 1992 (Willis, 1992). The feasibility of this approach was tested by Cornell and Yale Universities over a five year period (Chapman et al, 4), and both had slightly different approaches. Yale focused on digitizing microfilm, while Cornell focused on multiple formats for information on brittle paper, both with an eye for cost-effectiveness (Chapman et al, 6). These projects revealed that many variables factor into a hybrid project, including the characteristics of the materials being reformatted, the capabilities of available reformatting technologies, and what the final product will be used for (Chapman et al, 6). Clearly this is an area requiring further study to assist in the creation of cost-effective technologies and techniques of sufficient quality to preserve information, although this study does show that, despite their many problems, microfilm and digital formats can be used in a complimentary manner.

Bibliography

Arms, Caroline R (1996). Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress. *D-Lib Magazine*, April 2006. Accessed 3/31/2009 from http://www.dlib.org/dlib/april96/loc/04c-arms.html

Baker, Nicholson, "Deadline: The Author's Desperate Bid to Save America's Past," *New Yorker*, July 24, 2000, pp. 42-61.

Beagrie, Neil (2001). Going Digital: Issues in digitisation for public libraries. *Earl: The Consortium for Public Library Networking.* Accessed 4/22/2009 from http://www.ukoln.ac.uk/public/earl/issuepapers/digitisation.htm

Chapman, Stephen, Conway, Paul and Anne R. Kenney. Digital Imaging and Preservation Microfilm: The Future of the Hybrid Approach for the Preservation of Brittle Books. *Council on Library Resources.* Accessed 4/22/2009 from http://www.clir.org/pubs/archives/hybrid.pdf

Chapman, Stephen and Kenny, Anne R. (1996). Digital conversion of research library materials: a case for full informational capture. *D-lib Magazin*e, October, 1996. Accessed 4/22/2009 from http://www.dlib.org/dlib/october96/cornell/10chapman.html

Class notes for Topics: Resources/Services—Preservation (Instructor: Nancy Kraft). Notes by Julia Skinner, 3/23/2009.

International Federation of Library Associations and Institutions (1998). Digital Libraries: Definitions, Issues, and Challenges. *International Federation of Library Associations and Institutions*. Accessed 3/31/2009 from http://www.ifla.org/VI/5/op/udtop8/udtop8.htm

McMillan Library (2006). Digitization for Public Libraries. *McMillan Library.* Accessed 3/31/2009 from http://www.mcmillanlibrary.org/programs/digitize.html

Mindwrap, Inc. “OCR/Text Search.” *Mindwrap, Inc.* Accessed 5/10/2010 from http://www.mindwrap.com/infoblurbs/info_ocr.html

University of Georgia Libraries (2008). Georgia Newspaper Project. *University of Georgia.* Accessed 4/22/2009 from http://www.libs.uga.edu/gnp/faq.html

Willis, Don (1992). A Hybrid Systems Approach to Preservation of Printed Materials. *Commission on Preservation and Access.* Accessed 4/22/2009 from http://www.clir.org/pubs/reports/willis/index.html