

Editorial

Selection for digital preservation

Michael Seadle

The author

Michael Seadle is Editor of *Library Hi Tech*.

Keywords

Archives management, Digital storage, Copyright law

Abstract

This editorial discusses long-term archiving and long-term access to digital documents, with an emphasis on criteria for selection. Selecting materials for digital preservation depends on whether the materials are both valuable and endangered, whether appropriate digitization procedures and standards for these materials exist, and whether copyright allows reasonable access for educational and research purposes.

Electronic access

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0737-8831.htm

Recently I was asked to speak on a panel at the German Library Congress in Leipzig about “long term archiving and long-term access to digital documents”, with an emphasis on criteria for selection. This issue is less controversial than it was in 1992, when Anne Kenney and Lynne Personius wrote their landmark “Study in digital preservation”. Three of their principal conclusions were:

- (1) Digital image technology provides an alternative – of comparable quality and lower cost – to photocopying for preserving deteriorating library materials . . .
- (2) Subject to the resolution of certain problems, digital scanning technology offers a cost effective adjunct or alternative to microfilm preservation . . .
- (3) Digital technology has the potential to enhance access to library materials (Kenney and Personius, 1992).

Kenney and Personius were also clear that a number of problems remained, including the need to institutionalize technology refreshing. Not all of these problems have been solved, and a few new ones have been added to the list, such as establishing “authenticity” in a digital environment (Smith, 2000).

Nonetheless, digital preservation has grown increasingly acceptable as librarians begin to realize that digital preservation has nothing to do with transient media like tape, CDs or hard drives, but rather with the ability of systems to make perfect copies and with the foresight to have enough copies that no statistically plausible set of failures can eliminate them all. Projects like LOCKSS (Lots of Copies Keeps Stuff Safe) from Stanford University have gone a long way toward making digital preservation a reality (see Reich and Rosenthal, 2001). We have reached a point where the discussion has shifted from whether digital preservation is feasible to how to select appropriate materials.

Selecting materials for digital preservation depends on three criteria:

- (1) whether the materials are both valuable and endangered;
- (2) whether appropriate digitization procedures and standards for these materials exist; and
- (3) whether copyright allows reasonable access for educational and research purposes.

Valuable and endangered

The value of a work represents both an economic and intellectual calculation, and together these

Received 7 March 2004
Revised 14 March 2004
Accepted 19 March 2004



have a direct relationship to the work's survival chances: high overall value means a high probability of copies surviving. For example, each genuine Gutenberg Bible has a value in the millions of dollars. Few copies of the original artifact exist, but its intellectual contents are as safe as contemporary society can make them. Countless versions of the original text, revised texts, translated texts, even facsimile versions are available in paper, and now a digital version of the Göttingen copy is available (Lossau, 2000). A library fortunate enough to have an original can largely ignore it for long-term digital archiving, unless that copy has some previously undiscovered unique aspect, such as handwritten notes by Martin Luther, that no one else has captured.

Digital archiving could arguably reduce the economic value of an expensive original by increasing copies and making its contents much more widely available, though no evidence exists to suggest that this in fact is happening. If anything, access seems to expose artifactual originals to wider markets, increasing their economic value. Unless human cupidity changes, the economic value of digitized originals seems safe.

Low-value materials are, however, often genuinely endangered. The classic example is gray literature: those materials never published through commercial presses, not offered in bookstores, never registered with ISBN or ISSN or other schemes, and never inventoried except occasionally through the good fortune of archival organization. Gray literature composes large portions of the Web, and only Brewster Kahle's Internet Archive is making an earnest attempt to preserve it all[1]. Some efforts exist to reformat older paper-based gray literature. Michigan State University, for example, digitized its American Radicalism Collection in the late 1990s[2]. The collection includes fliers, pamphlets and drawings, often reproduced crudely on mimeograph machines or early photocopiers. Often the ink was already fading. They would not have survived another decade with the relatively heavy use they received from students and researchers who discovered an interest in the topic.

Between these extremes lies the bulk of printed materials. On average their economic value, as measured by used bookstores, hovers around \$5. Books with pictures may be worth more. Many works are sufficiently unsaleable not to be worth the cost of warehousing them, and end up in the dumpster. Their intellectual value may well exceed their economic value for a few scholars at some indefinite future time. These kinds of works represent the core of a good research library collection, which tries almost by definition to have broad or nearly comprehensive collections in a

least a few chosen subjects. Neglect protects them to some degree from the ravages of human oils, excess light, and exposure to the elements. Neglect also often condemns them to remote storage areas whose climate controls may lack sophistication. Works from the acid paper period will burn slowly away. They fall in the broad middle range for their survival chances as well as their economic value. A few libraries, notably the University of Michigan (University of Michigan Digital Library, n.d.), have started the systematic digitization of these materials using low-cost techniques that require the bindings to be chopped off, and the pages sheet fed in bulk through scanners. Long-term, these works are much more likely to survive in digital form than in paper, and their use will grow because Web-based discovery and access offers a vastly greater market for their intellectual value.

Standards and procedures

Many libraries think of digital archiving mainly in terms of reformatting existing paper materials. This is partly because early projects like Cornell's concentrated on printed materials, and thus developed standards for the digital versions using TIFF, SGML and now XML. Good training programs exist to ensure reasonably efficient procedures. Although some issues remain, long-term digital archiving of text-based materials seems reasonable.

The standards and methods for multimedia preservation remain far more volatile. Older multimedia that is currently stored in analog form on magnetic tape or on color film is more seriously endangered than all but the most acidic books. As with books, those items with particularly high economic value tend to have many more copies, but the inevitable loss in each generation of analog copies sets a threshold of quality, and even modern color film has some tendency to fade. The National Gallery of the Spoken Word project has helped to establish some standards (Seadle, 2004), including the use of WAV files, but issues like the preferred sampling rate remain in dispute with some arguing for 96kHz for all materials, and others defaulting to the more widespread 44.1kHz "CD audio" standard, which already captures more than the human ear can hear and has long been used for music.

Standards can hardly be said to exist for digital video. Most digital video is too highly compressed and too prone to loss when expanded to be considered reliable preservation, but the file sizes for storing uncompressed video make any large-scale effort economically infeasible. Falling disk storage prices may change that, however. An

additional complication is the fact that contemporary video often starts digital and incorporates software dependencies from the editing tools.

A larger version of the same problem exists for software preservation. Software is, of course, born digital, but it is also born with operating system or device dependencies that make the equivalent of reformatting just as necessary. Emulating old platforms and migrating software to run on new platforms are both alternatives, and at this point neither seems obviously preferable (Hedstrom and Lampe, 2001). Long-term archiving for software is very much an open issue.

Access

Digital preservation is no guarantee of access. US law allows libraries to create up to three digital copies of endangered works, but it is important to note that the law explicitly limits access to the premises:

The right of reproduction under this section applies to three copies or phonorecords of a published work duplicated solely for the purpose of replacement of a copy or phonorecord that is damaged, deteriorating, lost, or stolen, or if the existing format in which the work is stored has become obsolete, if:

(1) the library or archives has, after a reasonable effort, determined that an unused replacement cannot be obtained at a fair price; and

(2) any such copy or phonorecord that is reproduced in digital format is not made available to the public in that format outside the premises of the library or archives in lawful possession of such copy[3].

While some wishful arguments suggest that "premises" for a university library could include the whole campus, many who deal regularly with the law would not want to risk such a broad interpretation of a word that seems plainly to mean a single physical building.

Copyright restrictions are the chief reason why most digital archiving has concentrated on pre-20th century materials. Some exceptions to copyright protection exist, particularly within US law, and some institutions are willing to make a risk assessment that allows access to works whose rights owners seem unlikely to protest. Others, especially state-supported libraries, are more risk-averse, and spend significant resources trying to get permission to give access to materials they want to digitize.

Conclusion

For libraries that deal primarily with paper documents, selection for long-term digital archiving is limited mainly by choices about the value of the originals and the degree to which copyright law allows access. Standards and procedures are reasonably well established to give the works a good chance to long-term survival. Since most libraries fit this category, digitization projects can be expected to flourish.

For libraries with significant audio, multimedia, or software collections, as does Michigan State University's Vincent Voice Library, long-term digital archiving continues to have a significant research aspect. The copyright issues for multimedia especially can be particularly complex because of the potential for multiple ownership. Libraries selecting these kinds of materials for digitization need to be prepared for ongoing change.

Notes

- 1 See www.archive.org/
- 2 See <http://digital.lib.msu.edu/onlinecolls/collection.cfm?CID=1>
- 3 17 USC 108, United States Code, Title 17, Chapter 1, section 108, available at: www.copyright.gov/title17/92chap1.html#108

References

- Hedstrom, M. and Lampe, C. (2001), "Emulation vs migration: do users care?", *RLG DigiNews*, Vol. 5 No. 6, December, available at: www.rlg.org/preserv/diginews/diginews5-6.html#feature1
- Kenney, A. and Personius, L. (1992), "The Cornell/Xerox/Commission on Preservation and Access Joint Study in digital preservation", available at: <http://palimpsest.stanford.edu/byauth/kenney/joint/>
- Lossau, N. (2000), "Göttingen Gutenberg Bible goes digital", *D-LIB Magazine*, Vol. 6 No. 6, June, available at: www.dlib.org/dlib/june00/06contents.html (accessed March 2004).
- Reich, V. and Rosenthal, D. (2001), "LOCKSS: a permanent Web publishing and access system", *D-LIB Magazine*, Vol. 7 No. 6, June, available at: www.dlib.org/dlib/june01/reich/06reich.html (accessed March 2004).
- Seadle, M. (2004), "Sound preservation: from analog to digital", in Lynden, F.C. (Ed.), *Advances in Librarianship*, Vol. 27, April, Academic Press, New York, NY.
- Smith, A. (Ed.) (2000), *Authenticity in a Digital Environment*, Council on Library and Information Resources, Washington, DC, available at: www.clir.org/pubs/reports/pub92/contents.html (accessed March 2004).
- University of Michigan Digital Library (n.d.), "Making of America IV: the American voice, 1850-1877", available at: www.umdl.umich.edu/moa4/overview.html (accessed March 2004).