From Double Fold to Double Bind

Sarah Thomas
Carl A. Kroch University Librarian
Cornell University
201 Olin Library
Ithaca, New York 14853-5301
<set9@cornell.edu>

This paper is based on a presentation at the Digital Libraries Symposium, Elsevier Science, New Orleans, Louisiana, January 19, 2002.

Abstract

Academic libraries see digital preservation as part of their fundamental mission to guarantee enduring access to the record of civilization's accomplishments and discoveries. The nature of digital documents presents unique complications which challenge a library's ability to assure long-term availability. Publishers view electronic backfiles as economic assets, and libraries do not control the physical object. Agreements between publishers and libraries require reaching an understanding on technical issues, the degree of access to the data held in a digital repository, and on the financial responsibility for preservation. This paper describes some of the efforts underway in the United States to establish standards for repositories and to implement digital archiving for electronic journals.

The oldest object in the Cornell University Library is a cuneiform tablet dating from 2250 BCE. Apart from serving as an object of study by scholars of the Near East, the clay tablet is summoned forth from our vault as symbol of ancient civilizations when schoolchildren visit. The urge to preserve books, manuscripts, and other carriers of information has developed strongly in librarians both for the content they hold and for their intrinsic artifactual value. Librarians have proved to be faithful custodians of objected entrusted in them, and they have safeguarded their holdings against the ravages of war or nature or guarded them against destruction when the writings fell into political disfavor.

One of the hallmarks of the library today is its stability, particularly with regard to its collections. Unlike the bookstore, whose stock consists of in-print volumes and whose success depends on rapid turnover of inventory, the research library rarely deaccessions a unique item, and its readers reasonably expect to retrieve items previously consulted from week to week, year to year, and century to century. Librarians, especially academic librarians, consider a commitment to ongoing and enduring access a core value of their organizations, recognizing their role as stewards of cultural and intellectual heritage. Consequently, in the 1980s, when library materials manifested themselves vulnerable to decay, concerned librarians organized themselves to spread awareness of the brittle books problem, and they marshaled the support of the National Endowment of the Humanities and others to fund preservation reformatting.

Preservation today is an integral part of the library culture, although the commitment to an active preservation program is by no means a deeply imbedded aspect of all research libraries. It is primarily the largest academic libraries that devote significant resources to preservation, and even some among those institutions rely heavily on soft dollars. Active preservation is a Johnny-come-lately in the library scene, emerging as a significant functional responsibility only when the legacy of the combustion of nineteenth-century chemicals with wood pulp paper became recognized. In a recent Association of Research Libraries report on preservation, only about 80 of the 118 respondents to a preservation survey had a preservation program managed by a preservation administrator.[1] Enduring access to books has benefited from the widespread distribution and redundancy of holdings of titles as well as from the slowness of decay.

When digital technologies emerged and the ephemeral quality of electronic data became apparent, libraries immediately recognized the threat of loss. The Commission on Preservation and Access and RLG Task Force on Archiving of Digital Information, chaired by Don Waters and John Garrett, drew attention to the problem in 1995 and issued a set of guidelines and recommendations that remain valid today, six years after they published their report. However, the problem has remained knotty, complicated by several factors. Libraries had adopted their preservation responsibilities gradually, more as a by-product of their collection than by design. The frailty of paper revealed itself slowly; objects stood passively on shelves. By contrast, digital technologies supplanted one another in very short cycles, with a document's generational span lasting sometimes only months or a few years before they were seemingly irrevocably irretrievable. Second, the library, as owner of the container or paper publication, possessed a tangible object to preserve. Librarians did not distinguish between the carrier and its contents, and they wanted to provide appropriate stewardship for the artifact. Digital objects, by comparison, were often not "owned" in a permanent sense, but were rather "licensed," by libraries and consequently not even always accessible and hence not even physically in the library's custody. Furthermore, some librarians recognized preservation as an "unfunded mandate." By the 1980s, it was too late to negotiate with nineteenth-century publishers over the deteriorating quality of their stock. Those publishers had embraced a new printing technology that resulted in slow fires or a ticking bomb on the shelves of libraries. Today, perhaps librarians can learn from the past and strike a different deal with publishers. Yet, even if librarians want to shoulder the responsibility for preservation unilaterally, suddenly publishers are not so sure they want to entrust libraries with their data. Where paper inventories had been chiefly a costly liability for publishers, digital technology has turned backfiles into assets with an undetermined lifespan and replete with multiple repackaging opportunities. In the past, publishers often did not retain paper archives, as became evident when they needed to turn to libraries to fill in gaps of issues during digital conversion projects. In the digital age, the journal article has become a component of a larger database that has considerable ongoing economic potential, and the publisher often wants to retain complete control over a proprietary format.

The new medium has placed libraries and publishers in a double bind, especially where content appears in parallel paper and digital formats. Although the publisher often expresses the wish to streamline his operations and publish in a primary digital mode, displacing printing to the reader and eliminating the hassle of distribution, the library and the publisher remain locked into dual formats. One of the chief reasons cited by librarians for retaining paper is the inability of the publisher to guarantee the longevity of documents. This is both a matter of technical standards as well as the ability to demonstrate trustworthiness. Project Harvest, Cornell's Mellon-funded digital archiving project, conducted a study of agricultural libraries that found that 84% of libraries surveyed would cancel print if a reliable archive were built. Although the sample was too small to be statistically reliable, its results reinforce anecdotal evidence that endurance is a prerequisite for complete dependence on electronic resources. Yet until subscribers make a transition to a digital format, neither the publisher nor the librarian can realize the benefit of the digital-only model. Although the savings of either party cannot be fully calculated, some libraries have explored the possibilities. The Drexel University Library, without any pretension to a role as the preserver of literature, but comfortable in its role as consumer, documented savings by switching wherever possible to electronic as the preferred format, and by refusing to accept paper, even if it came bundled with the digital, and even if they had paid for it.[2] However, for most institutions today, there are still cultural reservations about moving to an entirely digital environment, since many readers are reluctant to relinquish the feel and familiar contextuality of paper. Kevin Guthrie has documented this conundrum in a JSTOR survey of faculty who noted, in response to the statement: "regardless of what happen with electronic journals, it will always be crucial for libraries to maintain hard copy archives," that this statement described their sentiments "very well" for 48% of them and "somewhat" for 30 %. At the same time, 97% of respondents (76% answering "very well" and 21% "somewhat") that with "more and more journals becoming available electronically, it is crucial that libraries, publishers or electronic databases archive, catalog, and protect these electronic journals." Although 77% *now* thought it was an important role of the library to serve as an archive, they anticipated a declining responsibility for the library in the future, with only 68% responding that in five years it would be an important aspect of the library.[3] Guthrie has also observed that although JSTOR is at its core a preservation enterprise that creates the opportunity to reduce redundant holdings and claims to shelf space, librarians continue to respond to it mainly as providing better and more convenient access to journals for faculty and students.[4] Although archiving was definitely seen by all parties (authors, readers, publishers, librarians) as extremely important, there was no consensus about who should be responsible for archiving digital documents or how it should be done. At issue is whether an archive, be it JSTOR or any other, is a service for publishers, a subjected-based access tool for scholars, or a management tool for libraries, enabling them to manage their physical assets and allocate the space in their facilities in a cost-effective way. It is the tension between the various, sometimes competing and conflicting functions that creates confusion between librarians and publishers in the establishment of digital repositories.

To move off this dead center, various organizations—the Council on Library and Information Resources, the Digital Library Federation, the Coalition for Networked Information, the National Science Foundation, the Society for Scholarly Publishing, and the Andrew W. Mellon Foundation—collaborated in a series of meetings where stakeholders rehearsed the technical, legal, economic, and social issues relating to digital preservation. From these meetings emerged gradual consensus about some aspects of the digital archives in the beginning of the 21st century. To ensure endurance, multiple agents should hold data in multiple locations. Since publishers can go out of business or merge with corporations for whom the bottom line may be of greater consequence than some superannuated publications of dubious revenue generation; governments may topple or impose a political agenda on a national library, and natural or manmade disasters may strike in any region, the parties supported distributed archival repositories. Also, because of technical vulnerability, the participants in these meetings often were proponents of different technical implementations. If one proved unstable, there would be a back-up elsewhere. There was considerable debate about what constituted a functional archive. Must a journal have all of its links and the full functionality of the publisher's interface in a true archive? Some publishers were relatively indifferent to the preservation of the full context of a journal, including information on its editorial board or advertisements, and thought only the articles themselves needed to be captured. Could one really know if all data were preserved if they were in a "dark" archive, an inaccessible virtual lockbox? Or would the archive have to be "lit" or accessible to ensure that it worked? What event would "trigger" access to the archive? The publisher taking the journal offline? The passage of time, measured in months or years? The transfer of a title from one publisher to another? What would it cost to support the archive, and what responsibilities did it entail? If a university entered into a contract to preserve the publisher's electronic journals, and that publisher went out of business, was it the library's role to provide ongoing access—at no charge?, or at what charge to the journal's subscribers? Did the archive them need to have a separate agreement with subscribers?

The issues are complex and varied. In *Attributes of a Trusted Repository: Meeting the Needs of Research Resources, an RLG-OCLC Report*, there is an imposing list of qualities and requirements for a digital archive:

- Auditability, security and communication;
- Compliance and conscientiousness;
- Certification, copy control, and following rules;
- Backup policies and avoiding, detecting and restoring lost/corrupted information;
- Reputation and performance;
- Agreements between creators and providers;
- Open sharing of information about what it is preserving and for whom;
- Balanced risk, benefit, and cost;
- Complementarity, cost-effectiveness, scalability, and confidence; and
- Evaluation of system components.[5]

The report proposes a framework that includes:

- Administrative responsibility;
- Organizational viability;
- Financial sustainability;

- Technological suitability;
- System security; and
- Procedural accountability.[6]

In the past year funding from the Andrew W. Mellon Foundation has provided the impetus for exploring most of these topics in greater depth. To generate some practical experience about the application of these abstract concepts, the Mellon Foundation awarded grants to seven institutions, including the New York Public Library and the university libraries of Cornell, Harvard, MIT, Pennsylvania, Stanford, and Yale. Updates on their progress in moving from plan to program in the area of digital archiving are posted on the Digital Library Federation home page.[7] In addition, these and other issues are discussed at length in Dale Flecker's excellent paper "Preserving Scholarly E-Journals."[8] Flecker notes the assumptions that informed the various Mellon archiving projects:

- "Archives should be independent of publishers, and that archiving needs to be the responsibility of institutions for whom it is a core mission;
- Archiving should be based on active partnership with publishers, and that it will require a different kind of license agreement than the normal content usage license;
- Archives should address preservation over very long timeframes (100 years or more); long timeframes are likely to raise issues very different from those encountered in daily service provision;
- Archives will need to conform to standards and best practice guidelines as they evolve in the digital world and should be subject to auditing and certification;
- Archives should be based on the Open Archival Information System reference model, currently being vetted by the International Organization for Standardization (ISO). This model is a careful analysis of the types of data and processes required for an archive to maintain content over extended timeframes."[9]

Cornell's particular interest stems from its vigorous preservation program and its substantial digital library initiatives. With over two terabytes of images in its Making of America project, the Core Historical Literature of Agriculture, digitized manuscripts, photographs, and visual images, the Library has, in effect, become a digital publisher. In the early days of electronic resources, Cornell did not always practice what it now preaches. Cornell's pioneering digital effort in the late eighties to convert core mathematical texts was threatened with extinction less than a decade later because the library had kept files on a server that was failing. "Rescuing," that is, extracting and writing to tape the approximately 750 math texts cost $30,000 in 1996, but providing access to them, that is, moving them from a dark archive to a brightly lit one, cost another $70,000. That cruel lesson sensitized the Library to both the need for archiving protection and its potential costs.

For the Mellon planning grant Cornell proposed a disciplinary-based approach since this approximated the real-life environment of the library, where holdings derived from multiple publishers. The subject-based repository, developed to its fullest extent, would create opportunities for access to a broad spectrum of related material. It addresses the interests of the scholar and researcher, rather than being publisher-centric. The disciplinary-based focus has made the planned implementation more complex, since there are multiple agreements to execute with different publishers and numerous formats to accommodate. However, Cornell had existing relationships with a number of publishers of agricultural titles, based on its TEEAL (The Essential Electronic Agricultural Library) project, which had negotiated the right to create a "Library in a box" of CD-ROMs of 140 agricultural journals to distribute for an affordable price in developing countries (see http://teeal.cornell.edu/). Cornell's agricultural and life sciences library also has an ongoing relationship with the U.S. Department of Agriculture (U.S.D.A.) to make available statistical reports from the Economic research Services and other U.S.D.A. units. The experience of loading and serving this data contribute to a very early awareness of the fragility of digital documents and led to Mann Library's organization of a conference on the importance of digital preservation in 1997. The Core Historical Literature of Agriculture project also provided another foundation block on which to rest the concept of the subject-based archive (see http://chla.library.cornell.edu/). Since 75% of core agricultural journals are now available in electronic format, a repository concentrating on agriculture seemed not only appropriate, but also even urgent. Cornell also brought to the table its ongoing research in digital preservation funded by the National Science Foundation as part of the DLI II program (see http://www.library.cornell.edu/preservation/prism.html). This effort, known as Project PRISM, attempts to characterize

> the nature of preservation risks in the Web environment, develops a risk management methodology for establishing a preservation monitoring and evaluation program, and leads to the creation of management tools and policies for virtual remote control. The approach will demonstrate ho Web crawlers and other automated tools and utilities can be used to identify and quantify risks; to implement appropriate and effective measures to prevent, mitigate, recover from damage to and loss of Web-based assets; and to support post-event remediation.[10]

Project Harvest, as the Mellon project is called, has uncovered a complex landscape. Anne Kenney and Nancy McGovern reported at the Coalition for Networked Information Task Force meeting in December 2001 on the work of Project Harvest to date (see http://www.library.cornell.edu/harvest/). An important milestone was a meeting in early September to lay the groundwork for the repository. Attending this meting were representatives of the American Dairy Science Association, Academic/Elsevier, the American Phytopathology Society, BioOne, CAB International, the Canadian National Research Centre, Wiley, the National Agricultural Library, and USAIN. Among the findings of this meeting were that there was still not agreement about who should do digital archiving. Although librarians trust third party archiving, as evidenced in an informal Project Harvest survey in which 90% of respondents preferred multiple custodians rather than a single party preserver, many publishers were insufficiently aware that others didn't trust them to archive materials responsibly or to be the sole custodian of their output. Many publishers were unaware of the requirements of reliable archives. Publishers and libraries were united, however, in their common distrust of government.

Perhaps the area of greatest debate has been that of the "lit" archive. Publishers are uneasy about permitting access, even carefully restricted access, to their assets. They can see the value of "lit" metadata, since that would link information seekers to the publisher articles and therefore have the possibility of generating revenue, but a fully accessible archive has scant appeal to them. The economic model is still murky, but what is emerging is an approach in which the publisher compensates the archiving party more if the archive is dark, and it is the publisher and his subscribers who are the primary beneficiaries. This approach might bundle a tax for preservation into the journal's subscription rates. Alternatively, publishers could create a perpetual care endowment at the point of deposit of a title. If the archiving agent receives access to the journal in lieu of a subscription, compensation would be correspondingly lowered. If a "lit" subject-based repository generated revenue, publishers would receive a share, and the potential expanded market might provide the incentive to allow access. Anne Kenney and Nancy McGovern, in a forthcoming paper for the Council on Library and Information Resources, posit that access to a market that includes lifelong learners and developing countries will expand revenue opportunities beyond the present base of academic libraries. They anticipate that the subject-based repository will bring economies of scale and the interoperable quality of the archive will enable the content to be mined by publishers and other value-adders to generate new products and revenues. Ideally, the content would be free, with profit based on services derived from manipulation of the content or supported through some other subsidy. And, despite the unease with which the government is regarded, its funds are welcome should it wish to designate them as a cultural and intellectual heritage responsibility. Flecker suggests that the funding and sustaining of archives will depend on a hybrid model incorporating many of these elements.[11]

While the amount that is known about archiving is relatively limited, there abide many misconceptions. One is that archiving is practically free, since the cost of storage has declined and promises not to be a major factor in the future. Archiving entails much more than disk space, however. Reliable systems require tending; formats ingested change and require reprogramming; file back-ups require monitoring; and equipment must be housed and upgraded regularly. Some people have also suggested that since libraries regard preservation as a core activity, they should fund it themselves, rather than looking to publishers to support third-party archiving. Yet, as we have seen, libraries cannot independently archive digital materials which they do not own. Others have inferred that the savings libraries will realize once they get out of the double bind of maintaining paper and digital will more than compensate for the investment of a few institutions in digital archiving. But how is the archiving institution to capture the savings from hundreds of beneficiaries? Institutions must consider the extent to which they wish to expend their local resources for the greater good of society.

CONCLUSION

It is important that we all remember that archiving is both very important and complex. Furthermore, the technology of archiving is still evolving; preservation of digital documents will require multiple, distributed repositories with diverse technological implications; and the costs of archiving are unknown. In such circumstances, publishers, librarians, and other stakeholders (e.g., authors, scholars, societies, and government agencies) need to work together to create successful models to guide digital archiving and preservation. Initiatives underway through the funding of The Andrew W. Mellon Foundation, the National Science Foundation, and the Library of Congress should point the way in the United States for successful models.

NOTES AND REFERENCES

[1] *ARL Preservation Statistics, 1997-1998* (Washington, D.C. : Association of Research Libraries, 2001), p. 5.

[2] Carol Hansen Montgomery, "Measuring the Impact of an Electronic Journal Collection on Library Costs: A Framework and Preliminary Observations," *D-Lib Magazine* 6 (10) (October 2000).

[3] Kevin M. Guthrie, "What Do Faculty Think of Electronic Resources," *ALA Annual Conference Participants' Meeting*. June 17, 2001 (Available: http://www.jstor.org/about/faculty.survey.ppt) (accessed February 4, 2002).

[4] Kevin M. Guthrie, "Archiving in the Digital Age," *EDUCAUSE Review 36(6)* (November/December 2001): 56-62.

[5] RLG-OCLC, *Attributes of a Trusted Repository: Meeting the Needs of Research Resources,* A Report (draft for public comment) (Mountain View: RLG, 2001), p. 12.

[6] Ibid., pp. 11-12.

[7] See http://www.diglib.org/preserve/ejp.htm (accessed February 4, 2002).

[8] Dale Flecker, "Preserving Scholarly E-Journals," *D-Lib Magazine7* (9) (September 2001).

[9] Ibid., p. 3.

[10] Anne R. Kenney, Nancy Y. McGovern, Peter Botticelli, Rich Entlich, Carl Lagoze, & Sandy Payette, "Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project PRISM," *D-Lib Magazine, 8 (1)* (January 2002).

[11] Flecker, "Preserving Scholarly E-Journals."