By Dale Flecker

Digital Archiving: What Is Involved?

Advances in technology have brought many useful tools to scholars, teachers, students, and others, but these advances have also brought significant threats. Digital information is not durable, and in the rush toward technological innovation, many in higher education have ignored the need for digital archiving and preservation. Only recently has attention to preservation been increasing in library and scholarly circles. Among the first kinds of resources to receive such attention are electronic journals, which have grown dramatically in number and importance. Two years ago the Andrew W. Mellon Foundation awarded grants to seven research libraries to study e-journal archiving. The thoughts below are based on Harvard University's study in the Mellon program.[1]

Archiving and preservation do not just happen. Digital materials are surprisingly fragile. Their viability depends on technologies that rapidly and continually change. Many valuable paper resources that have been acquired by individuals or organizations and stored in little-visited recesses remain viable decades later. That will not happen with digital materials. There is no digital equivalent to that decades-old pile of *National Geographic* magazines in the attic. Changes in technology will ensure that over relatively short periods of time, both the media and the technical format of old digital materials will become unusable. Keeping digital resources usable by future generations requires conscious effort and continual investment.

E-journals require a fundamentally different model. The business and service model for e-journals differs from the model for paper journals in two significant ways:

1. Paper resources are physically controlled by libraries, for which long-term preservation is a fundamental role. Library subscriptions for e-journals, on the other hand, provide access, not copies. In general, e-journals are held only in the systems of publishers or their agents, for whom archiving for future scholars is not a core mission.
2. Generally, copies of paper journals and books are widely replicated. This redundancy is an important safeguard, helping to ensure that some copy of any given work will persist over time. Most electronic publishers maintain multiple copies in mirroring systems, but these provide only a partial form of redundancy. Such systems do little to protect against economic failures, changes in institutional policy, conscious "amendments" (like the changes in the Stalin era, when those who had fallen from grace were removed from Soviet photographs), systematic software errors, and the like. Effective redundancy should prevent destruction by any single element or agency.

Ownership matters. When an institution buys a copy of a physical work, it owns that copy, and many rights follow. For e-journals and other materials accessed over the Internet, rights to make copies, alter formats, and take other actions attendant to archiving are not granted in normal use licenses. Archiving programs will require explicit agreements with copyright owners and will be shaped by conditions these owners impose. Agreements with different communities of owners will involve different issues:

- *Fear of competition.* Publishers are concerned about who can access archived content and under what circumstances. Many librarians believe that archival content must be used to ensure its viability, but such use can represent competition to owners.
- *Lack of interest.* Publishers for whom libraries represent a significant market will likely respond to overtures from archiving institutions. Other publishers (both mass market and niche) are likely to be removed from the concerns of scholars and the issues of archiving.
- *Lack of relationship.* Many Internet resources are provided without explicit license, and archiving institutions have no direct relationship with the owners. Locating owners and getting their attention and cooperation will be challenging.

What gets archived? Archiving efforts of any kind are selective. In the e-journal realm, one naturally thinks first of selection by title: archives preserve titles that fall within their collecting scope. The question of scope, however, is more subtle. With each new source of archival content, libraries must negotiate licenses and business arrangements, work out the interoperation of systems, and maintain relationships. It may be economically more efficient for an archive to handle all the

e-journal output of any given source, thus selecting by producer rather than title. The second level of selection applies *within* each journal title. E-journals are complex objects, and various components will pose archiving challenges:

- Most e-journals treat content such as statements of copyright, ownership, and editorial policies, and names of editors and editorial boards, in reasonably unsophisticated ways outside of their content-management systems, and they change such data independently of issue publication. Archiving such content will differ technically and operationally from archiving articles.
- An increasing number of e-journals now include "associated datasets"—files of information such as research data, models, software programs, and audio or video supplementary to the article text. There is generally little control over the format or documentation of such files, making archiving problematic.
- Web links between references in e-journal articles and the referenced materials are also increasing rapidly. These highly dynamic links are frequently maintained by publishers in databases independent of the article containing the link. Acquiring such links and keeping them valid will be difficult.

In what way is material preserved? The phrase *digital preservation* is highly ambiguous, covering a spectrum of actions and results, including the following:

- Preserving only electronic bits, so that in the future, someone with appropriate skills and resources can figure out how to render the object or otherwise make it useable in then-current technology
- Maintaining usability as technology changes, so that users can view or use objects originally created in technology that has become obsolete
- Maintaining not just usability but also the original interface

Who should archive? Who should pay? In research libraries, there has been little distinction between archiving paper journals and providing access to paper journals. The copy of a journal used day-to-day was also the archival copy. Because many of the same activities contributed equally to service and preservation (e.g., proper housing, repair of damage, binding, reproduction of unusable copies), true preservation costs were hard to segregate. With e-journals, there will frequently be a distinction between the publisher's daily service and the archiving of content over time. Costs of archiving will become painfully visible. Particularly in the United States, where institutional funding is decentralized and there is little tradition of either top-down or coordinated planning, the issues of who should be responsible for archiving and of who should pay are difficult to resolve. Given the scale of the preservation challenge, it is highly unlikely that a single institution such as the Library of Congress or a single funding source such as the federal government will bear the entire burden. Only a small number of institutions need archive any given content to serve the requirements of all scholars, but the temptation to "let the other guy do it" will be great. Archiving is vulnerable to the free-rider syndrome. The scholarly community will have to create new understandings, and possibly new organizational structures, for archiving shared research resources.

What can be done now? Both the scale and the complexity of the digital archiving problem are daunting. There is great danger that those institutions that should act in this realm will be held back. An understanding of the technical, organizational, and financial aspects of digital archiving must be reached quickly. In the short term, the higher education community must take at least two initiatives:

1. Establish model archives to demonstrate feasibility and explore technical, structural, legal, and economic issues
2. Prepare the ground with key decision-makers to build an understanding in institutions, particularly in government and colleges/universities, of the nature and importance of digital archiving; archiving of the scale needed to support scholarship will require that these institutions assume responsibility and provide funding at a significant level

Related to these initiatives is a recent development of potentially great importance. The Library of Congress has received funding from Congress to plan a National Digital Information Infrastructure and Preservation Program of distributed digital archiving. The higher education community must monitor, encourage, and when possible, participate in this innovative program.

Illustration by Steve McCracken, © 2003

Note

1. The seven are the New York Public Library and the university libraries of Cornell, Harvard, MIT, Pennsylvania, Stanford, and Yale. Full results of the Harvard study and other studies can be found at <http://diglib.org/preserve/ejp.htm>.

Dale Flecker is Associate Director for Planning and Systems, Harvard University Library.