Preserving Scholarly E-Journals

Dale Flecker
Associate Director for Planning and Systems
Harvard University Library
dale_flecker@harvard.edu

Introduction

For research libraries, the long-term preservation of digital collections may well be the most important issue in digital libraries. In certain ways, digital materials are incredibly fragile, dependent for their continued utility upon technologies that undergo rapid and continual change. In the world of physical research materials, a great number of valuable research resources have been saved passively: acquired by individuals or organizations and stored in little-visited recesses. These physical resources are still viable decades later. This is not the case with the digital equivalents. Changes in computing technology will insure that, over relatively short timeframes, both the media and the technical format of old digital materials will become unusable. Keeping digital resources accessible for use by future generations will require conscious effort and continual investment.

North American research libraries have been discussing the issues of digital preservation for some years now, but the number of active programs remains extremely small. As the range of materials in collections available only in digital form grows, it will be increasingly important that libraries move from discussion to action. This paper introduces one initiative in the area of digital preservation and discusses some of the difficult issues it is raising.

E-Journal archiving

In the past two or three years, e-journals have become the largest and fastest growing segment of the digital collections for most libraries. Collections that a few years ago numbered in the few hundreds of titles now number in the thousands, and the rate of growth continues to increase.

In many ways, archiving and preserving e-journals will be dramatically different from what has been done for paper-based journals. In the paper era, there was large-scale redundancy in the storage of journals. Many different institutions collected the same titles. The copies of journals being saved for future generations were the same copies being read by the current generation of users. Many of the things that helped maintain journals for the long term (binding, repair, sound handling and shelving practices, environmental control, reformatting when usability was threatened) were not differentiated from what a library did to provide current services. Other than in the case of preservation microfilming and the odd instance of shared book storage facilities, there was little conscious coordination of preservation
activities, and in fact a level of redundancy was expected and thought useful.

The common service model for e-journals is quite different than that for paper journals. Most e-journal access is through a single delivery system maintained either by the publisher or its agent. There is little replication, and only a few institutions actually hold copies of journals locally. Libraries can fulfill their current service requirements without facing the issues involved in the preservation of the resources. Further, in the digital realm the issues involved in day-to-day service are quite different from those involved in long-term preservation.

The issue of long-term archiving and preservation of e-journal content has become one of increasing importance. Specifically because of archiving concerns, many research libraries continue to collect paper copies at the same time they pay for access to the electronic versions. This dual expense is not likely to be sustainable over time. Publishers are finding that authors, editors, scholarly societies, and libraries frequently resist moving to electronic-only publication because of concern that long-term preservation and access to the electronic version is uncertain. Of perhaps even greater long-term concern, while libraries continue to rely on the paper copy as the archival version, from the viewpoint of publishers it is increasingly the electronic versions of titles that are the version of record, containing content not available in the print version.

These tensions and concerns led to a series of meetings over the past few years among publishers, librarians, and technologists sponsored by a variety of organizations, including the Society of Scholarly Publishers, the National Science Foundation, the Council on Library and Information Resources, and the Coalition for Networked Information. While these meetings helped to identify many of the issues, they did not result in any specific follow-up action. Finally, in the summer of 2000, the Andrew W. Mellon Foundation, working with the Council on Library and Information Resources (CLIR), took the initiative to move beyond exchanges of viewpoint to experimentation and implementation.

Mellon/CLIR initiative

In a series of meetings with libraries and publishers, CLIR defined a framework for e-journal archiving [1]. Based on this framework, the Mellon Foundation then invited a number of research libraries to apply for one-year planning grants to develop projects to create and operate experimental e-journal archives. In December 2000, six planning grants were awarded, and a seventh grant was given for a related technical development. The planning grants took three different approaches:

- Three projects are publisher-oriented:
  - Harvard University proposed working with John Wiley and Sons, Blackwell Publishing, and the University of Chicago Press;
  - The University of Pennsylvania proposed working with Oxford University Press and Cambridge University Press;
  - Yale University proposed working with Elsevier Science.

- Two projects are subject-oriented:
  - Cornell University in agriculture;
  - The New York Public Library in performing arts.

- The Massachusetts Institute of Technology proposed investigating the challenging area of "dynamic e-journals" (scholarly web sites that aim to share discoveries and
insights, but do not feel bound by the conventions of "issues" and "articles" that have become standard in print).

A seventh grant was made to Stanford University to fund the further development and beta testing of the LOCKSS system, which is intended to automatically, and with little cost or overhead, support the large-scale replication of e-journal content [2].

The planning projects generally shared a number of key assumptions:

- 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.

Issues

Another key assumption of the Mellon initiative is that there will be relatively few archives holding any given set of e-journals, and that institutions operating archives will be doing so not just for their own users, but for the general community of subscribers and readers. In this environment, the design and operation of an archive therefore will be of concern not just to the publisher and the operating institution but to all who will rely on the archive. There are many important technical and policy issues raised by archiving projects, and it is critical that these be discussed by the general community of libraries, publishers, and scholars, and not just left to the archive and publishers involved.

Harvard and its publisher partners have been discussing and thinking about archiving for a number of months. Some of the more important questions identified so far in this process include those below.

What is the publisher/archive/subscriber relationship? Publishers and subscribers, and publishers and archives, have formal contractual relationships; does there need to be a formal relationship between archives and subscriber? (See Figure 1.)
Is archive content usually "dark"? "Dark" content is that which is not accessible for normal daily use. An archive that keeps its content dark poses less of a threat of competition to the publishers with whom it is working. A dark archive will also be relieved from having to maintain a current user interface, with all of the bells and whistles that users have come to expect, and from the complex task of maintaining information on who has access to what content. On the other hand, ensuring that content that is never used remains sound and free from degradation will be challenging.

When can archived content be accessed? If archived content is initially kept dark after deposit, under what conditions can it subsequently be accessed? Many archiving discussions revolve around the concept of "trigger" events, that is, conditions that change the access rule of the archive, for example:

- when a given journal is no longer accessible on-line;
- after a certain amount of time has passed since initial publication (this is the current policy of PubMed Central, which calls for deposited content to be openly available no more than one year after publication [3]);
- when a title changes hands.

Who can access archived content? If a trigger event happens, who gets access? Just subscribers (individual or institutional)? Controlling access in this way is complex. Keeping records of who has the right to access what and implementing appropriate access control mechanisms that recognize differential rights to various archived objects would be a major operational challenge.

What content is archived? At first hearing, most people assume that e-journal archiving is basically concerned with the content of journal articles. Indeed, while articles are the intellectual core of journals, in fact e-journals contain many other kinds of materials. Some examples of commonly found content are:
Which of these content types need to be archived and preserved for the future? Some of these types of materials will pose issues for publishers. Not all of these items are controlled in publishers' asset management systems. Some are treated as ephemeral, "masthead" information and are simply handled as website content. When such information changes, the site is updated and earlier information is lost. For example, few if any e-journals provide a list of who was on the editorial board for an issue published a year or two ago. Another difficult content type is advertisements. Advertisements are, of course, frequently not tied to any given issue, and they change over time with the business arrangements of the publisher. In some cases, advertisements are specific to certain populations, and what advertisements you see depend on who or where you are. (For instance, drug ads are frequently regulated at the national level.) Deciding what of all that is seen on e-journal sites today should be archived and maintained will require careful consideration by archives, publishers, and scholars.

Should content be normalized? The variety of formats of digital objects in an archive will affect the cost and complexity of operation. In order to control such complexity and cost, an archive might want to normalize deposited objects into a set of preferred formats whenever possible. Such normalization can happen at two levels:

- File formats. An archive might prefer to store all raster images in TIFF, for instance, and convert JPEG or GIF images into that format. Controlling the number of file formats will reduce the complexity of format monitoring and migration.

- Document formats. Many e-journal publishers encode article content in SGML or XML (or plan to do so soon). Most publishers create their own DTD (or modify an existing DTD) to suit their specific needs and delivery platforms. An archive might choose to normalize all such marked-up documents into a common DTD, reducing the complexity of documentation, migration, and interface software. (As part of Harvard's planning project, a consultant is examining the feasibility of creating an "archival e-journal DTD," which would be a preferred format for article deposit.)

Normalization and translation always involve the risk of information loss. In archiving there may well be a difficult trade-off between information loss and reduced complexity and cost of operation.

Should a standardized ingest format be developed? The OAIS model uses the concepts of "information packages," that is, bundles of data objects and metadata about the objects that are the unit of deposit, storage, and distribution by an archive. The model allows transformations to be done as objects move from one type of package to another. (See Figure 2.)
If as expected any given publisher is depositing content into a number of different archives, and any given archive is accepting deposits from a number of different publishers, standardizing the format of "Submission Information Packages" may reduce operational cost and complexity for both communities (although at the cost of devising and maintaining such a standard).

**Preserve usable objects, or just bits?** A key element in digital preservation is maintaining the usability of digital objects in current delivery technology as the technical environment changes over time. This process is usually assumed to be one of "format migration," that is, the transformation of objects from obsolete to current formats, although it could also be carried out through emulation, that is, maintaining current programs capable of emulating older technology, thus rendering obsolete formats. Whatever the method, the cost of preservation will be sensitive to the number and types of formats in an archive.

E-journals can contain a very wide range of technical formats, particularly as they begin to accept digital files created during the process of research (statistical datasets, instrument produced datasets, visualizations, models, video and audio files) that help validate, supplement, or further explain the basic content of articles. Whether it will be practical for archives to maintain current usability for such a diverse range of formats is far from clear. It is possible that archives will need to differentiate between formats where usability will be maintained and formats for which the archive will insure that the bits are maintained as deposited and that whatever documentation exists about them is kept useable for future "digital archeologists".

**Who pays what?** Archiving and preserving e-journal content will cost money. How much money is uncertain, and that is one of the many things the Mellon initiative will help clarify. Perhaps the most important single financial issue is how archiving can be implemented to minimize the cost to the community. The question of who pays is likely to be quite sensitive to the magnitude of the cost.
It is unlikely that archiving will be funded through a single source or even a single funding model. Over time, many different parties can be expected to contribute to the effort. Certainly for one-time or episodic expenses (systems implementation or re-implementation, large-scale format migrations, etc.), sources such as foundations or government funding programs are likely sources of support. But archiving is a continual process, with expenses incurred on an on-going basis. So, a reasonably secure, continuous funding source is required.

It has been suggested by some that archives could support themselves through fees to users for access. However, if the purpose of an archive is to provide failsafe access rather than daily service, this model will not provide on-going operational funding. Archiving is a form of insurance, and as with insurance one experiences the expense on an on-going basis but experiences the benefit only occasionally. Expecting to recover the cost of archiving only at the point at which access to the archived content is necessary is impractical. Such incidents will be rare and randomly timed, whereas the cost of archiving will occur from the first day of deposit.

Another widely-discussed option is for archiving to be funded by governments through the agency of national libraries or similar bodies, particularly for materials subject to copyright deposit. This may indeed be a good model for support of some archiving, but it is unlikely to be sufficient. Not every country will be equipped or financially willing to assume archiving responsibilities. More importantly, one archival agency for a work is insufficient. While digital archiving need not involve the scale of redundancy that we had in the paper era, some redundancy is highly desirable. There is too great a danger that a single incident, decision, or mistake can destroy what has been archived. A sound archiving model should involve multiple archival copies in the hands of different organizations, subject to different national laws and political influences, and dependent upon different technical infrastructures and preservation activities. The need for redundancy suggests that archiving cannot be left solely to copyright libraries and national funding.

For archives lacking independent funding such as governments can provide, the most attractive funding model may well be one that involves the deposit of funds to maintain materials at the same time the materials themselves are deposited. Such a "dowry" would insure the growth of funds in proportion to the growth in responsibility. Dowry funding might come through the agency of the publisher, but its ultimate source is most likely to be the subscribers to the archived journals. This funding could be made visible through the means of an "archiving surcharge" added to subscription fees, or it could be simply wrapped into the budgets of publishers and/or journal owners (such as scholarly societies). The centralized e-journal delivery model pursued by many publishers does not account for the core function of archiving, and introducing it back into the model can be viewed as a natural cost element of electronic publishing. Of course, the general scholarly community is the key beneficiary of archiving. Subscribers and scholarly societies can serve as surrogates for that community in the realm of e-journals, and a cost model based on funding from such sources is not unreasonable.

**Conclusion**

E-journal archiving has no easy analog in our current environment and raises many new issues requiring careful analysis and wide discussion. At least for the North American library community, the current Mellon-sponsored planning projects, and the likely follow-on projects discussed below, provide an opportunity to begin such thinking and discussion.

**Next steps**
The six current Mellon grants are intended to provide one group of libraries and publishers an opportunity to consider what would be involved in a large-scale e-journal archiving project. Topics under consideration in the planning year include:

- The rights and responsibilities of archives and publishers;
- The nature of the license under which an archive would have access to a publisher's content, including the key issues of who can access archived content and under what circumstances;
- Technical issues concerning the form and format of archival submissions;
- The technical architecture of an archive and the magnitude of development effort required to build one;
- Organizational models, operating characteristics, and on-going expenses of an archive.

This planning process will continue until early in 2002. The Mellon Foundation will then entertain proposals for follow-on projects to actually construct operating archives and operate them for several years. Up to four continuing projects may be funded. The intent of these follow-on projects is to accumulate sufficient experience with the operation and costs of archiving to help the scholarly community consider the most appropriate ways to institutionalize the critical function of preserving the scholarly record as it migrates to purely digital form.

Notes