# Planning for the digital archive: the Harvard e-journal experience

*Marilyn Geller*

*Preservation of material in library collections is a fundamental charge of libraries, and electronic resources are rapidly becoming significant components of these collections. Along with five other major U.S. research libraries, Harvard University is currently engaged in an e-journal archive design project under the auspices of a one-year planning grant from the Andrew Mellon Foundation. This article discusses many of the issues Harvard and its publishing partners are exploring in an effort to develop and refine the design of this archive.*

*Marilyn Geller, MSLS*
*Information Management Consultant, 436 School Street*
*Belmont, MA 02478 USA*
*Tel: 1 617-484-7379*
*Fax: 1 617-484-2629*
*E-mail: marilyn.geller@mind-spring.com*

Preserving the historical record of scholarly content is a fundamental charge of academic libraries. For a variety of media commonly held in libraries, producers and consumers have worked together to make preservation a reachable goal. The very best example of this collaboration is the development of standards for the permanence of paper that allow publishers to produce and libraries to store the printed historical record with a high level of confidence that this paper with its intellectual content will be available to scholars far into the future. The development of electronic resources that are appropriate for library collections has given publishers, librarians and service intermediaries many challenges, such as, licensing, providing access and integrating with an existing collection. To date, however, there has been much discussion but little real progress in terms of preservation of these resources. The same high level of confidence, which librarians have for the preservation of paper, must be present for electronic resources in order for these pieces of scholarly content to become truly a sustainable part of the library collection.

To that end, in the fall of 2000, the Andrew Mellon Foundation working with the Council on Library and Information Resources invited a number of U.S. research libraries to submit proposals for one-year planning grants to develop projects for the creation and operation of e-journal archives. E-journals are, of course, only one of the many types of electronic resources currently available, but they are a good starting place. There are now significant numbers of titles from a variety of sources; every major publisher now produces journals that are available via the Internet. Many of these titles have been published for several years making it reasonable to examine the archiving issue over time. And while it is hard to imagine that this format has stabilized, patterns have emerged that allow us to examine the

basic model and its permutation within the archiving context.

**Publisher-focused archive**

In December 2000, Harvard University was one of six major research libraries to receive one of these Mellon grants[1]. Harvard proposed to explore the development of an archive based on the collection of e-journals from specific publishers. There are, in fact, a number of different ways that an archival collection could be focused. In opting to work with specific publishers, Harvard intends to test the assumption that there will be some economies of scale in processing large numbers of titles from the same source. Underlying this assumption is the notion that each time a title is accepted into the archive, some preliminary work would have to be done to accommodate the format in which it will be received, and some human intervention will occur to establish appropriate transfer procedures. Further, we assume that one publisher will be able to deliver all journals in the same format and by the same process, thus allowing the archival system to ingest large volumes of material from one source with limited effort. It would, however, be unwise to test this assumption with only one publisher. To that end, Harvard has chosen to explore partnerships with three publishers, two commercial and one non-profit. During the period covered by the planning year grant, Harvard is actively working with Blackwell (including Blackwell Publishers, Blackwell Science and Munksgaard), John Wiley and Sons, and the University of Chicago Press.

Early on in the planning, the decision was made that the goal of this archive is to preserve the e-journal, not simply e-journal articles, which raises the question of what are elements that comprise an e-journal. In analyzing the content of e-journals, a list of the kinds of content that are part of the archiveable e-journal was drawn up. This list includes many elements that were deemed to be in the scope of the archive, such as: the articles themselves, supplementary files associated with the article (datasets, sound and video files, etc.), internal and external links, abstracts, table of contents, editorials, correspondence, reviews, editorial boards, copyright statements, editorial policy, reviewers lists, and threaded discussions. Some of the kinds of material that could be deferred, at least initially, include: information for authors, advertisements, and reprint, subscription and customer service information. Surprising for its obviousness is the editorial board. While editorial board changes are amended in the immutable print journal and, therefore, preserved, that is not always the case in the electronic environment, where editorial board information is often changed without regard to preserving its earlier iterations. In discussions with our publishing partners, we are beginning to sort out what is available, in what formats and whether it might be appropriate for inclusion in the e-journal archive.

**OAIS and METS**

At the same time that we are thinking about what will be deposited, we are also considering how it will be deposited in the archive. In the technical aspects of planning the e-journal archive, there is a key assumption for all of the Mellon grant projects that the archives would be based on the Open Archive Information System (OAIS) reference model[2], which describes deposit, storage and distribution of archived objects. According to the OAIS model, packages of information that include objects and metadata about the objects are ingested into the archive in the form of a Submission Information Package (SIP). Harvard is exploring the possibility that all SIP metadata will be encoded in XML files. It would be dramatically labor intensive to ingest packages constructed differently by different publishers, whereas standardizing the SIP format among publishers could potentially lessen these operational costs. Harvard is investigating using METS[3], a new XML Schema for encoding structural and related metadata. To explore the value of standardization further, Harvard has contracted with Inera Incorporated, a provider of SGML-related consulting services. In the early fall of 2001, Inera will examine the article DTDs of Harvard's publishing partners and others and identify the feasibility of constructing a common archival article XML DTD. In identifying the lowest common denominator, we expect that there will be some loss of information, and this will be measured against the potential gains that could be made in processing.

**Access control**

The issue of who has access to the archive, under what circumstances and in what manner, is under consideration. Options run the gamut from a completely dark archive, wherein no one has access for normal daily use, to a completely light archive, wherein everyone has access. Not unexpectedly, finding the right balance here is no small task. In order to ensure that the archived material does not degrade over time, it must be exercised. At the start, Harvard proposed that the archive should initially be semi-dark, permitting access only to Harvard's and participating publishers' archiving staff, to Harvard University Library's authorized users, through an online process, and to any user legitimately authorized by the publisher, through a batch transfer process, thus allowing for maintenance, auditing and minimal exercising of the data. The publishing partners had some concerns about this position that will collaboratively move the discussion forward. Their concerns include the preference for having users access their own embellished systems in lieu of Harvard's more stark archive, the concern for monitoring to guard against unauthorized use, and the reluctance to allow Harvard users to access material that Harvard has archived but not subscribed to or licensed.

The concept of trigger events, or circumstances under which archived material would move from a dark or semi-dark state to a light state, was originally proposed by one of the Mellon grant libraries. Examples of some trigger events might be: when the title is no longer available online from the publisher; when it is no longer available anywhere online; when the title ceases to be published, or after a defined amount of time has passed. It might be that these trigger events could be defined on a title-by-title basis or across an entire collection. Questions surrounding these trigger events might include the permanence of a trigger event decision (Can a "sunrise" decision be changed?), use of the copyright expiration of 75 years as the trigger, and the restricting of trigger events to include only two options: when the publisher releases a title or when the publisher is no longer available to be asked. Clearly, there is much to explore in this area.

**Long-term preservation**

In thinking about the issues of providing long-term access, one can hardly ignore the more fundamental issue of long-term preservation. While other options exist, migration, the transformation of material from an earlier format to another format compatible with whatever the current technology allows, seems to be the best way forward. In the best of all possible worlds, an archive would preserve the usability of all materials. However, the tremendously wide range of file formats may suggest a limit on an archive's ability to carry out this task completely. For those formats for which a clear migration path can be made, the Harvard archive would maintain the usability of data. For those formats outside the migration option, Harvard suggests that it would maintain the bits, that is, the archive would preserve material in its original format along with the metadata and technical documentation that would allow future technically sophisticated users to recover the intellectual context.

The numbers of files, and of file types that an archive can ingest and migrate, are only two aspects of the cost analysis involved in the archive. The development, administration and continued maintenance of the system as well as, to a lesser extent, the size of the archived content are also to be accounted for. By using Harvard's current digital library infrastructure, some costs can be shared. The remainder of the costs can be divided into one-time start up costs and ongoing costs. In any event, archiving will incur a significant expense, and while Harvard does not presume to recoup all costs associated with this archiving project, at least a portion of the financing should be shared among the community of beneficiaries, including, possibly, institutional subscribers and societies, for which journals are published as the key representatives of the scholarly community. We continue to pursue this idea trying to refine how the fee might be determined on a per journal basis and how such a fund might be collected and managed to ensure its availability for ongoing costs.

**Stakeholder rights**

In the course of discussing the financing of the archive, it was Harvard's publishing partners

who raised the issue of the fuller role of this community of beneficiaries. If stakeholders are asked to finance a portion of this archive, what are the other rights and responsibilities of this group and how do they become informed and make known their informed opinion on the direction of the archive? It is Harvard's current position that the university owns the archive and that this ownership allows the archive to use Harvard's digital library infrastructure. The publishing partners have suggested a range of possibilities for a stakeholder's group that run the gamut from a specific advisory group to a broader governing body that might oversee a loose coalition of archives. This is an ongoing area of discussion.

The one-year planning grant, which allows Harvard to develop a blueprint for a sustainable archive should be seen in the larger archiving context. This is only one of six planning projects by major research libraries that the Mellon Foundation is funding. The redundancy of digital archives is absolutely essential although it is most likely not necessary to match the redundancy of the print archives. Multiple archives that pursue different categories of collection and different technologies and that replicate data will only act to ensure that the scholarly record is preserved. Beyond that and in light of recent political events, it becomes clear that archives also need to be replicated in a variety of geo-political units.

Early in 2002, Harvard and the other grant recipients will submit their reports to the Mellon Foundation. As of this writing, Harvard intends to submit a follow-up proposal for a four to five year grant to actually build the functional archive in partnership with those publishers, with whom we can come to agreement and who have helped shaped this archival vision. While this will surely not be the last word in archiving e-journals, it will be a significant step forward.

**References**

1 For proposals and updates on all of these Mellon grants, see: http://www.diglib.org/preserve/ejp.htm

2 *Reference Model for an Open Archival Information System* (OAIS), CCSDS 650.0-R-1.1 (Red Book). Oxford, UK: Consultative Committee for Space Data Systems, April 20, 2001 <http://ssdoo.gsfc.nasa.gov/ nost/isoas/us20/ 650x0_010510.pdf>.

3 *Metadata Encoding & Transmission Standard (METS)*, Washington, DC: Network Development and MARC Standards Office, Library of Congress, 2001 <http://www.loc.gov/ standards/mets/>.