Report on the Planning Year Grant for the Design of an E-journal Archive

Presented by:
Harvard University Library Mellon Project Steering Committee
Harvard University Library Mellon Project Technical Team

To:
The Andrew W. Mellon Foundation
1 April 2002

Table of Contents

1. Introduction

Early in 2000, the Digital Library Federation, the Council on Library and Information Resources, and the Coalition for Networked Information sponsored a series of meetings with librarians, publishers, and licensing specialists to identify minimum requirements for e-journal archival repositories.[1] Based on a request from the Andrew W. Mellon Foundation to build on these requirements, the Harvard University Library was one of several research libraries that submitted a proposal for the design and planning of an electronic journal archive and subsequently received a one-year planning grant in December 2000. 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 intended to test the assumption that there would be some economies of scale in processing large numbers of titles from the same source. Between January 2001 and March 2002, the Project Steering Committee and the Project Technical Team (see Appendix A) worked together and with other Mellon grant recipients and publishing partners to identify needs and solutions.

2. Project Objectives

During 2001, Harvard University Library used its one-year planning grant for an electronic journal archive from the Mellon Foundation to explore and define both the business and technical issues of content, format and deposit mechanisms, access control and interface requirements, long-term preservation guidelines, costs of development, operation and maintenance of the working archive, and financial and governance models for a sustainable archive. The remainder of this report represents our research findings and current thinking on the design of a publisher based e-journal archive.

2.1 Archive Mission

Archives serve a variety of different functions in the larger society and even within the smaller scholarly community. Research libraries in particular serve to "support education, continuous learning and research"[2] for their designated constituents. The focal points of this type of collection are intellectual artifacts generally in textual and graphic formats. An increasingly significant amount of the intellectual content is published and distributed in electronic journals. This Archive's specific mission is to:

> *Preserve the significant intellectual content of a defined set of electronic journals independent of the form in which that content was originally delivered in order to assure that this content will be accessible to the scholarly community for the indefinite future in a readable format.*

2.2 Scope of this Project

Harvard has proposed to begin collaboration with selected publishers to build and archive each publishing partner's entire collection of e-journals that can be deposited according to agreed specifications. Moving forward, Harvard has envisioned working with multiple publishers to build an operational model archive and a large collection of archived e-journal content.

Functionally, the Archive is designed to render text and still images and other formats as practically possible with no significant loss in intellectual content. The Archive reserves the right to freely manipulate the internal format of the manifestation over time as long as the plain meaning of the intellectual content is preserved. In general, archiving takes place at a semantic level, not a syntactic one.[3] This allows the Archive to be constructed around the principle of data format migration, rather than access system emulation.

2.3 Publishing Partners

Initially, Harvard proposed to select potential publishing partners who produce a significant volume of content in digital format to test the scalability of the ingest process and the Archive. Based on the stated criteria for the grant, a key characteristic of any publishing partner would be a strong interest in archiving and a willingness to invest time and resources in the project. Beyond that, it was assumed that any publishing partner would have to possess a high level of technical expertise in order to contribute to the technical planning process. We recognize that this assumption is not appropriate when dealing with smaller or less technologically sophisticated publishers. However, we are optimistic that much of the development work planned for this Archive will produce sharable tools and infrastructure that may be adaptable for other publishing environments and assume that archives will, when dealing with less willing or able partners, themselves assume more of the responsibility for technical integration. How these different models for publisher-archive interaction will change the economics and operation of archives needs exploration.

For the purposes of moving forward in exploring business and technical aspects of the Archive, Harvard held preliminary discussions with Blackwell Science and the University of Chicago Press. John Wiley was considered as a possible partner if Harvard could come to agreement on an electronic journal license; this was subsequently completed and Wiley was included in the group of potential partners. The Massachusetts Medical Society, publisher of *The New England Journal of Medicine* (NEJM), declined to participate in further discussions, citing concern about the time and labor involved in view of other commitments and the lack of perceived need for an archival partner. Our discussions with Blackwell Science were expanded to include both Blackwell Science and Blackwell Publishers which later merged to form Blackwell Publishing. The resulting group of three potential partners has given Harvard the opportunity to explore a variety of issues from the perspective of a large privately-held commercial organization, a large

publicly-held commercial organization, and a small non-profit organization, each of whom works closely with scholarly societies. Collectively, these publishers produce 1,137 journals in electronic format. While the ultimate goal of our discussions with these partners was to come to an agreement on the business and technical aspects of the design for the e-journal Archive, the intermediate goal of understanding the issues from a variety of perspectives proved to be immensely valuable.

2.4 Content

Harvard proposed to build a publisher-based archive that would hold the entire collection of e-journals offered by selected publishing partners. The underlying assumption is that material published by any given publisher is based on a common production process and that this uniformity will make it easier for the publisher to supply standardized input to an archive, thus simplifying the Archive's ingestion process. In addition, by working in depth with publisher partners, the Archive can create a more sophisticated archiving plan informed by an understanding of the publishers' internal systems and specific content.

2.4.1 Issue-centric Focus

Currently, the general practice of our three publishing partners is to regard their electronic titles as parallel (and possibly supplemented) manifestations of their print editions. Although all three have indicated a willingness to explore the modality of issue-less publishing at some point in the future, none is actively experimenting with the concept. Since the notion of issue-less journals represents a major shift in serials publishing practice and is unsupported by current library systems and scholarly practice of citation, we have decided to retain the concept of the issue as central to the design and implementation of the Archive. However, we retain some flexibility in this matter by allowing a loose definition of issue as a publisher-specified aggregation of content items without necessary regard for fixed publication patterns. As will be apparent in Section 4, this issue-centric focus is central to the design and implementation of the Archive. From the perspective of ingestion, this issue-centric focus allows Harvard to control receipt of content and to determine by examining the sequence if something has not been received. Ingest will be based on our publishing partners depositing new e-journal content in the Archive on a predefined schedule. The suggested schedule, each issue to be deposited prior to release of the next issue, will allow for an even and manageable ingest flow.

2.4.2 E-journal Components

Many think of e-journal archiving only in terms of preserving articles. However, e-journals actually contain a complex range of components. The target for preservation is defined in this project as the electronic version of the journal. At this early experimental stage, it was deemed best to preserve as much as possible in the categories of components and functionality. To determine what components and functionality of e-journals exist, twenty-one journals were examined (see Appendix B). This sample included eleven titles from the publishing partners and ten titles from other sources. They covered a wide range

of disciplines and represented titles available in print and electronic formats as well as titles only available electronically.

For all journals with printed versions, information about both the printed and electronic versions is available in the electronic version. All journal sites have basic descriptive information such as scope and purpose, subject coverage, and copyright statement. Most sites also have ISSN, frequency statement, indexing and abstracting service coverage, current editorial board, submission information, subscription and reprint information, and contact information. This category of information is equivalent to the front matter of the printed journal and provides an essential intellectual infrastructure for the journal. In discussion with the publishing partners, it has become evident that much of this front matter is not preserved in the electronic environment: only the most current version is available. While the editorial board information is associated with a particular issue in print versions of journals, this is not the case in the electronic version. Linking the appropriate version of front matter to specific issues is now an important item for the publishing partners to explore.

Within the issue or discrete publication bundle, all journals have table of contents information. Items listed in the table of contents include articles, case reports, comments, communications, correspondence, responses, dialogue, columns, editorials, letters to the editor, book reviews, conference notes, news, announcements, interviews, errata, volume indexes, subject indexes, membership lists, and reviewers.

Advertisements are found in two of the journals reviewed. Advertisements present a particular challenge for archiving. First, it is not uncommon for Web advertisements to be served from an organization other than the content publisher so that archiving agreements would not necessarily cover their deposit. Second, advertisements are frequently "dynamic," changing from day to day. The same page viewed on different days can have different advertisements. The advertisement seen in one country may be different from that seen in the same context in another; drug advertisements, for instance, are regulated at the national level and therefore vary with the country of receipt. What is the appropriate advertisement to archive with a given issue? When would dynamic advertisements be archived?

Web advertisements will be an important source for documenting contemporary business, society, design, and technology. However, they represent a minor type of content for scholarly e-journals. Harvard has decided not to archive advertisements as part of this e-journal archiving project. We hope, however, that someone somewhere is archiving Web advertisements more generally as part of the documentation of our time.

For journals examined in this survey, most articles include an abstract in HTML. Generally, the articles are delivered in both HTML and PDF; however, other formats noted include Postscript, TeX, and DVI, delivered as individual files or as aggregations bundled together in ZIP packages. The HTML versions of article content offer thumbnail images of tables, generally in GIF format and occasionally in JPEG format. Tables are

also included in the HTML versions. Figures, equations, symbols, and other graphics are delivered in GIF and JPEG formats.

One of the great powers of digital journal articles is that they are not limited to linear text and static pictures. Increasingly, articles include "supplementary materials," digital files of many types. These files can include digital materials used in the research described (statistical or instrumentation datasets, for example) or materials that expand on or illustrate topics discussed in the article (simulations, or tables too large for inclusion in the base article, for instance). These supplementary files represent a significant resource but also a significant challenge to the Archive.

In general, there is little control over the technical formats for supplementary files, no guidance to authors about good practices in the creation of such files, and little editorial analysis of the file content. The technical heterogeneity of these materials could introduce a wide and ever-growing range of formats into the Archive, significantly increasing the complexity of the preservation task. The lack of guidelines and quality control means that unlike the case of articles themselves, the Archive is faced with objects of unknown virtue and potentially troublesome content.

One of our publishing partners suggested that it would be very useful to publishers if archives could provide guidance on preferred technical formats and practices that are well suited for preservation and archiving. In the current environment, there is nothing to suggest to authors how to create digital objects with a greater chance of long-term viability. Many authors and their editors are very concerned about the longevity of their publications and if given guidance, may well be willing to change practices in ways that will reduce the complexity of preserving supplementary article content.

Most articles offer internal linking among components within the HTML version of the full text. External links are also common and link out to authors' email addresses, author-provided URLs, citations in external indexing resources, other articles by the same author, and related articles. Linking is a fast-changing area in e-journals and represents one of the great value-added features of electronic over paper publishing. However, links pose significant challenges in archiving. The largest application of links today is for references, allowing a user to navigate automatically from one article to a related cited article. Most publishers, however, do not simply insert static links in references. Most links today are in the form of Digital Object Identifiers (DOIs), a type of persistent link or "name" that remains valid even if the cited work moves between systems or publishers. Frequently publishers keep a database of references and only determine the DOI for a citation — a costly process — once. This allows the DOI to be reused each time the same work is cited. Further, the number of DOIs for retrospective articles is growing very rapidly so that even if no link were available for a reference when an article is originally published, one could be added later when the DOI becomes available.

While the majority of links found in e-journals today are for references, there is every reason to expect linking to become richer with time. There is already significant growth

in links to "knowledge bases" such as GenBank.[4] As these new types of links occur they will be added by automated means to existing articles.

Because links are dynamic and are expected to grow with time for already archived articles, it is unclear whether an archive should attempt to capture those links available when an article is ingested. Would it be better for archives to implement the types of dynamic linking systems that publishers use, allowing for ever more rich links for archived content? Or should archives arrange for publishers to periodically resend the links for articles submitted earlier? In either case, there will be complexity involved in supporting links for archives, but links are a vital type of content and users will likely be dissatisfied if they are not included in archived content. This is an area requiring more exploration. All journal Web sites include browsing functionality and most include search capability. All journals except one include help features.

2.4.3 User Survey

As we considered which components of electronic journals should be archived and how much of the look and feel of an e-journal should be preserved, there were concerns that costs would prohibit comprehensive archiving. It is clear that all articles, reports, columns, editorials, communications, abstracts, errata, and correspondence must be archived. It is less clear which other components should be preserved and how nontraditional components — such as links, threaded discussions, data sets, and data simulations — should be handled. To address these questions, we designed a survey that we completed by interviewing faculty in the sciences (see Appendix C). The survey focused on journal functionality including browsing, searching, image size and content (including cover images), tables of contents, subject and author indexes, advertisements, editorial board membership, editorial policy, announcements, membership lists (for societies), reviewer lists, copyright, guidelines for authors, career/job information, and business information (advertising guidelines, subscription information, and contact information). Faculty were primarily concerned with the reliable archiving of scientific content, specifically articles, reports, editorials, and other original content, plus functionality, including browsing, searching, and printing. Hierarchical links among volumes, issues, tables of contents, and articles were identified as important. Threaded discussions were of interest, but not considered critical by some faculty since they are not peer-reviewed. Access to original data sets provided by the authors was also considered useful, although providing reliable and accurate links to materials not maintained by publishers is problematic.

2.4.4 Components in Scope

Based on analysis of the e-journal sample and the user survey, Harvard has defined a preliminary list of materials deemed to be in the scope of archival collection and those currently out of scope of the archival collection, and will work with selected publishers to identify which components are available. Deposits will include not only journal articles, but also associated materials (e.g., references, external links, abstracts), author-created supplementary digital files (e.g., datasets, sound files, simulations), other editorial journal

content (e.g., editorials, reviews, communications, letters, threaded discussions), and selected masthead information (e.g., editor, editorial board, copyright statement). Materials currently defined as in scope should be deposited, while those defined as out of scope are not expected but may be deposited if available, with the exception of advertising which will not be accepted.

The following components should be deposited in the Archive:

Articles: This includes the text and auxiliary files such as (but not limited to) graphics, figures, tables, and/or photographs that constitute the "article proper."

Supplementary material/enhanced contents: When the author has deposited to the publishers digital objects related to the article such as (but not limited to) datasets, sound or video files, and/or computer programs — as opposed to pointing to such resources at alternate sites — those materials will be included in the Archive deposit.

Author supplied references

Links to external resources

Abstracts

Tables of contents

Placeholder files for non-deposited objects[5]

Other editorial content: This includes (but is not limited to) research, reports, columns, editorials, communications, correspondence, reviews, letters to the editor, and commentaries.

Bibliographic descriptions: This includes formatted metadata describing articles and other editorial content.

Editorial boards

Editors

Threaded discussions

Copyright statements and information

Editorial policies

Reviewer lists

Journal descriptions

Cover images from the corresponding print issues

2.4.5 Components Currently Out of Scope (Not Deposited)

The following types of components are not expected or required for deposit in the Archive:

Information for authors: This includes copyright transfer agreements and guidelines for manuscript preparation and submission.

Subscription information

Advertisements

Other business information: This includes reprint ordering information, information for posting advertisements, contact information, and customer service information.

Additional information: This includes career/job information, etc.

3. Business Model

3.1 Access Issues

One of the criteria to be met by this archive design is to make preserved information available to libraries under conditions negotiated with the publisher. Policies governing access to the Archive must address three questions: who can access the Archive, under what circumstances, and how will access be obtained. From our earliest discussions about access, it was determined that publishers should deposit materials into an initially dark archive as these materials become available. A dark archive is one that allows no access for routine scholarly use. As a result of some event — a "trigger" event — material in the Archive would be made available to some set of scholarly users resulting in a light archive.

Once the Archive has accepted preservation responsibility for deposited material, that content is then subject to periodic auditing to insure the efficacy of the Archive's preservation regimen and of the working of the repository system. Auditing is particularly important for the Archive since it represents the only use of archived content while that content is in its initial dark period prior to the occurrence of a trigger event. The composition of the auditors could be drawn from domain experts, subject area librarians, faculty, and scholarly societies. It remains an open question whether domain expertise is required or practical. However, initial quality control and internal and external auditing are not sufficient to insure the viability of files over time. During the grant year, Harvard had informal discussions with several organizations that gather content from various sources and store that content. Although initial quality control procedures varied, each organization maintains that actual usage is one of the best mechanisms for insuring content viability. The issue an archive must face is whether dark content not used regularly by expert users is an adequate and reliable preservation model.

3.1.1 Authorized Users

Harvard originally proposed that the Archive should initially be semi-dark, permitting access only to Harvard University Library's authorized users through an online process and to any user legitimately authorized by the publisher through a batch export process. Such access would allow for maintenance, auditing, and minimal exercising of the data. The publishing partners had some concerns about this position including:

- the preference for having "real" users access their own embellished systems rather than Harvard's more basic Archive interface;
- the need for monitoring to guard against unauthorized use;
- the reluctance to allow Harvard users to access material that Harvard has archived but not subscribed to or licensed.

During the resulting discussions, Harvard agreed that throughout the dark period the Archive would be accessible only by its operators and by a designated outside auditing authority. Under certain circumstances, archived material might be made available to authorized users who can present proof of their legitimate right to specific materials.

The Harvard University Library Access Management Service (AMS) allows fine granularity of access control, down to the level of permitting access to a particular object by a particular user through the use of a particular application. This level of control is appropriate and practical to support auditing, but expensive to extend to large-scale external use. An archive intending to restrict access to only those who have had a past subscription to the archived content would bear considerable expense to gather the required licensee information and to build and operate the appropriate access management system. Rather than pursue this option, Harvard believed that the increasingly wide-spread adoption of the "moving wall" concept (in which content becomes publicly available after a specified time period) in the scholarly journal environment suggested a more practical approach. After deposit, archived materials would enter an initial "dark" period chronologically bounded by the trigger events defined in the submission agreement. When any one of the trigger events has occurred, material would be accessible without restriction.

3.1.2 Trigger Events

At various times over the planning period, the following possible trigger events (conditions which would cause archived content to become publicly available) were discussed:

1. When material is no longer accessible online from the publisher. This trigger was intended to support the essential "failsafe" function of the Archive, insuring continued access to the scholarly record. After much discussion the provision was modified in several ways. "Material" was replaced with the more specific "volume or time-based unit of the title," recognizing that portions of the electronic run of a

journal might have different availability over time. The new wording allows for part of a run to become accessible from an archive when it is no longer available elsewhere. "Accessible online" was modified to "accessible online either from the publisher or from another source as a discrete title." This allowed titles that were transferred from one publisher to another to remain "dark" if other triggers had not occurred. It also protected access to the content through title and issue units, as opposed to having the content buried in an undifferentiated aggregate database. The final form of this trigger was thus: "when a volume or time-based unit of the title is no longer available online either from the publisher or from another source as a discrete title."

2. When the publisher sells or otherwise transfers the rights to publish a given title to another body. Publishers rightfully objected that this meant that titles could not be sold, as the "lightened" content in an archive would greatly reduce the value of the retrospective content. This trigger was dropped in later discussions.

3. When the material has been in the Archive for "n" years ("n" being a time period to be agreed to by Harvard and the publisher on a title-by-title basis). This trigger occasioned the greatest amount of discussion and was not fully resolved during the planning period. It was refined in later discussions slightly to "after a defined amount of time of the publisher's choosing has passed, to be determined by title and volume or time-based unit."

4. When the title ceases to be published. Some publishers objected that a ceased title may still have residual economic value. The provision was dropped from later discussions.

5. When the content enters the public domain.

Trigger events are one of the key provisions of an archiving plan. They define when the preserved content, which someone has made a considerable investment to archive, is useable. Because they touch on areas that affect the commercial value of the content and thus on the publisher's income, publishers are legitimately quite concerned that they be carefully constructed. All parties to archiving — authors, publishers, archives, subscribing libraries — have an interest in the details of trigger definition. This is an area requiring further discussion among the concerned parties.

3.2 Economic Issues

Economic issues are paramount in planning for archiving. In the paper environment, many of the costs of archiving are buried in library budgets. Much of what is done to preserve print journals is the same activity needed to provide day to day access to the literature. In the e-journal environment we are moving to an architecture which separates archiving from daily service, making archiving costs painfully apparent. Further, it is unlikely we will have or need the same large-scale redundancy in e-journal archiving that we have had for paper journals. It is more likely that the operating costs of archiving will be centered in only a few places, raising obvious issues of how to spread costs fairly.

Understandably, economic issues were discussed extensively during the planning year: within Harvard, with our publisher partners, and with other institutions thinking about archiving.

We do not know what archiving will cost. Beginning to understand real costs will be one of the key objectives of any implementation project. It is clear, however, that keeping costs low is enormously important; the magnitude of costs will greatly influence the outcome of the question of who is willing to share in the cost of archiving. Harvard identified some strategies for controlling costs. First, by building an archiving program over a larger digital library environment, activities (preservation monitoring), organization (computer operations), and technical infrastructure (a digital repository) already in place can be used to support the archiving activity. Second, applying smart automation to the process of adding content to the Archive can be used to reduce labor, and thus the cost, of ongoing Archive operation. Third, limiting the functionality of the Archive allows us to eliminate costly development components such as subscription management systems.

Even if an archive is successful in controlling costs, the ultimate question remains "who pays." Some options for support of archiving are inappropriate for Harvard's Archive. The institution will not simply bear the cost of archiving on its own as a national library might. Certainly Harvard would expect to contribute to the cost of archiving, but it will be difficult to convince university administrators that the institution should simply dedicate resources at this level for the common good.

Some have suggested an archive should support itself, at least in part, by providing services others would pay to use based on the archived content. We have not pursued this option for several reasons. First, some of the publishers we have talked to were initially unwilling to allow the Archive to resell services using their content. Second, making the Archive dependent on building marketable services adds a major new dimension to the already large task of archiving. The information marketplace is full of smart and aggressive players. Competing in this marketplace requires both capital and a sophisticated understanding of many complex markets — there is not likely to be a single product design that suits different topical domains. It is far from clear that all domains will provide opportunities for the development of profitable services to support archiving. Lastly, archiving must be a perpetual activity. Funding for a sustainable archive cannot be dependent on a service that may or may not be viable in some future marketplace.

Harvard has proposed that funding for a sustainable archive accompany the deposit of content from the outset. The model initially proposed to our partners was that there be an explicit "archiving surcharge" publishers would charge all institutional subscribers to archived titles that would be passed on to the Archive. The intent of the proposal was that the community that benefits from the Archive also assumes some share of the cost of archiving. Our publisher partners did not want the Archive to dictate pricing policy to them, so this model was modified as follows. The publishers would pay the Archive an annual fee for archiving. They in turn could collect the required funds from any one of a number of sources including authors (through page charges), sponsoring scholarly

societies, or subscribers as appropriate in individual cases. The archiving fee would be composed of two elements:

1. An "ingestion" fee to pay the operating costs of day to day receipt, quality control, and archival preparation of new content.

2. An amount to be added to the Archive endowment to cover the long-term cost of storage and preservation activity. Endowment is a very appealing model to pay for a long-term commitment such as archiving.

In designing this Archive we understand the experimental nature of this project as a means of developing sustainable models and encouraging more work in the field. As with all experiments, it is quite possible that new choices and better alternatives may arise out of this work. For this reason, it is important to establish an agreed upon exit strategy. Harvard has suggested that if it chooses to cease archiving any given set of materials, this specific content of the Archive would be transferred to another archive selected by agreement between Harvard and a stakeholder community. In addition, an amount of the remaining archiving fund proportional to the amount deposited by each publisher for those e-journal titles would be transferred to the new archiving organization. In the case where a publisher chooses to terminate its relationship with Harvard's Archive, materials that have previously been deposited will remain in the Archive, and deposit of all titles will continue until the current volumes are complete.

4. Technical Model

4.1 Technical Infrastructure

The software and hardware environment for the Archive will rely upon existing technical infrastructure developed by the Harvard University Library over the past three years under the aegis of its Library Digital Initiative.[6] The core component of this infrastructure is the Digital Repository Service (DRS), an Oracle-based repository for digital objects. Within the DRS, object content streams are stored along with their associated administrative and structural metadata. The DRS, now in its second production release, currently maintains over 240,000 objects with a total size of 120 gigabytes. The DRS is responsible only for the managed preservation of the objects deposited within it; resource discovery and delivery are handled through independent systems.

Digital objects are delivered out of the DRS through media type-specific delivery applications. Delivery applications are available for *simple objects*, those atomically composed of a single physical content stream, such as a raster image file; and *complex objects*, logical aggregations of intellectually or structurally related content streams, such as an electronic monograph structurally delivered in a page turning navigational environment. Additional applications are under development for streaming audio and video media types.

Dependent upon the access rules defined for a particular digital object, delivery applications may make use of the facilities of the Access Management System (AMS) including user authentication, profile, and authorization.

Digital objects stored in the DRS can be given persistent identifiers registered with and resolved by the Name Resolution Service (NRS). NRS identifiers and their resolution mechanism are compatible with IETF recommendations for URNs.[7] The NRS is composed of two subsystems: an Oracle-based administrative system that maintains the mappings between URNs and URLs, and a THTTP-based resolution server.[8] Archived e-journal components will be named in the NRS at a level of granularity corresponding to that of discovery and delivery, that is, at the issue and item level.

Descriptive metadata useful for resource discovery is contained in catalog systems external to the DRS. For the purposes of the Archive, title and issue-level descriptive metadata will be stored in HOLLIS, Harvard's Integrated Library System (ILS), searchable through a Web-accessible OPAC. Title-level descriptive information, such as ISSN, publisher, etc., will be captured in MARC bibliographic records, while individual issue-level information, such as chronology, enumeration, etc., will be stored in related holdings records. Issue-level catalog metadata will also include an actionable link in the form of an NRS persistent identifier to a dynamically generated, issue-specific Web page, providing table-of-contents-like access to individual e-journal issue items. Item-level metadata will be managed in a new item-level catalog, implemented either as a separate database in the ILS, or as a stand-alone XML database. It will provide a mechanism for search and browsing and will include actionable links to dynamically generated Web pages displaying individual journal items.

4.2 Archive Architecture

The design of the Archive was conceived in relation to the OAIS reference model, with its six main archival functions: ingest, data management, archival storage, preservation, access, and administration (see Appendix D).[9]

Large portions of the operational aspects of the Archive are amenable to automation, including areas such as publisher registration and profiling, Submission Information Package (SIP) submission, ingest validation at a syntactic level, SIP-to-AIP (Archival Information Package) transformation, archival storage deposit, preservation migration, routine reporting, and handling and responding to access requests. This degree of automation is achievable through a strict requirement of publisher compliance to formal standards for SIPs and the definition of a small set of normative data formats. The resulting uniformity of the Archive input stream and the canonical nature of internal archive storage practices provides the opportunity to rely upon automated systems to perform routine ingest, archival storage, data management, and access functions. Through the collaborative development of community standards, there is good potential for the sharing of common infrastructure components between archiving projects and institutions.

Dependent upon the implementation details of a new ILS currently undergoing installation, serial check-in and claiming operations may also be amenable to automation. The primary remaining tasks that will require manual intervention include ingest validation at a semantic level (the degree to which this is feasible is subject to further investigation); preservation planning, primarily the monitoring of the technical obsolescence of data formats; and ongoing periodic auditing of archived materials.

4.2.1 Ingest

The Ingest function is responsible for accepting and acting upon submissions of material for deposit into the Archive. The technical infrastructure necessary to support the following subtasks within this function are for the most part not currently extant, and their implementation would occupy the majority of the first year of an implementation project.

4.2.1.1 SIP

From the point of view of a content provider the Archive is opaque with a single defined input interface, the SIP. The structural envelope of the SIP in the Archive is provided by METS (Metadata Encoding & Transmission Standard), a comprehensive XML framework for encapsulating digital objects.[10] The unit of submission to the Archive is the e-journal issue. Physically, the SIP will take the form of a three-level file system hierarchy corresponding to the e-journal title, issue, and items. The title-level directory is empty and is present only to provide a common structural parent. The issue-level directory contains a METS file encapsulating all issue-level metadata and pointers to issue-level content files (e.g., masthead, editorial board, cover image) and to the item-level directories, each of which in turn has a METS file containing all item-level metadata and pointers to item-level content. Content object technical metadata stored in the METS files provides the necessary representation information to facilitate archival preservation activities and content delivery. A preliminary draft specification of the SIP is undergoing public review and comment.[11]

One of the core concepts of the Archive is the use of a common archival item-level schema for articles and "article-like" content. The implications of and many design principles for such a schema are defined in a preliminary feasibility study on the subject, commissioned by Harvard and authored by Inera.[12] The design of this schema will begin with an investigation of existing common schemas, such as the ISO 12083 and the PubMedCentral Document Type Definitions (DTDs), for possible use as is or as the basis for additional development. If a new schema should be necessary, in whole or in part, the design and documentation of it may be contracted out to an appropriate consultant and developed with coordinated input from the larger community. As will be set out in the Archive submission agreement, when practical it will be the responsibility of content providers to transform their journal content from its internal native form into compliance with the Archive's schema when the publisher has content marked up in SGML or XML. In order to transform content to this schema, participating publishers will need a significant level of internal technical expertise or access to external technical expertise

and the resources to implement the transformation workflow. It is clear that not all publishers of scholarly content will have these assets. We will attempt to provide whatever technical assistance is feasible towards this effort including documentation, tools, or training for those publishers who are in a position to work with the Archive. All three of our publishing partners have agreed in principal to the use of a common archival schema. The potential for archive simplification due to this type of normalization emphasizes the importance of collaborative work within the archiving community to achieve consensus on common standards.

Since digital preservation activities are performed on a type-specific basis, minimizing the number of acceptable data formats can reduce the complexity and cost of archive operations. To place this on a formal basis, the Archive will define a small set of preferred normative formats. In general, a single normative format will be defined for each functional category of content: for example, XML for metadata, XML for full-text (using numeric character references for non-ASCII Unicode characters, named character entities for non-Unicode characters, and MathML for mathematics), TIFF for raster still images, XML/SVG for vector still images, etc. Data submitted in non-normative formats will be transformed upon ingest into an analogous normative format whenever possible without significant loss of intellectual meaning. For example, submitted JPEG files will be transformed into their analogous TIFF representations. Content objects submitted to the Archive in non-normative formats that are not susceptible to transformation into a normative analogue will be accepted, but only under the proviso that they may be preserved only at the bit level — i.e., in the form of the initially deposited bit stream — and that their usefulness over time may become problematic.

It is our belief that long-term archival preservation requires the initial capture of intellectual content at the highest possible resolution, finest possible granularity, and most abstract representation. Additional criteria for the selection of normative formats include open standards, mature and robust technology, long-term viability, prevalence of commercial grade tools, and the potential for instantiated data objects to be created as far upstream as possible in publishers' production processes. The composition of the set of normative formats will undergo periodic review to insure they remain appropriate for archival purposes with regard to continual technological advances.

During our initial evaluation of the PDF format with regard to its inclusion in the set of archival normative formats, several undesirable characteristics of PDF were discovered. Foremost, perhaps, is the fact that PDF is a proprietary rather than an open standard. Although Adobe has published the specifications, this is a matter of company policy and subject to unpredictable change. The built-in extensibility of PDF allows it to provide a structural envelope into which content can be placed in a variety of base formats. For example, PDF files can be composed partially or entirely of raster page images rather than the actual text. Internal PDF content streams can be formatted or compressed using completely private schemas, some or all of which may be resistant to archival preservation. Also, PDF is most generally encoded in a binary rather than an ASCII form which tends to increase the complexity if not the difficulty of processing. Many of the

challenges to be faced in preserving PDF content are examined in detail by John Ockerbloom in a recent paper in *RLG DigiNews*.[13]

However, despite our reservations concerning the long-term preservability of PDF, the fact remains that it has found overwhelming utilization in electronic publishing. As the Archive grows over time to encompass publishers beyond our initial partners, we anticipate that in a non-trivial number of cases PDF will be the only content format that some publishers will be able to provide to the Archive. Thus, we will include PDF as a normative format. However, we will attempt to constrain the specific internal format of PDF content through a published set of best practices (full-text, rather than page images; standard, rather than private compression; no encryption, etc.) to which publishers will be strongly encouraged to conform. We will store these PDF versions that publishers will deposit and will also use them as part of the quality assurance effort.

All relevant information concerning the various data formats recognized by the Archive, both normative and non-normative, will be stored in a central format registry. Depending upon the use for particular pieces of format information, it may be encoded in human or machine-readable formats. These data will include items such as formal format name, version history, pointer to authoritative specification, name of maintenance organization, MIME type, technical metadata schemas, compliant tools, and validation and migration processes. Since format-specific expertise is widely distributed in the archiving community, as is the need for the information captured, the format registry represents another instance of a common infrastructure piece that is deserving of community-wide development and maintenance.

4.2.1.2 Submission Session

The Submission Session refers to the operational process of physically transferring a SIP from a content provider to the Archive. Due to the potentially large number and size of e-journal issue components, we will investigate mechanisms for implementing the submission process with regard to the granularity of the transfer (i.e., a single aggregated unit versus individual file components); fixed (e.g., DVD) versus electronic medium of submission; and in the case of the later, limited throughput of commercial network connections and the reliability of standard protocols.

4.2.1.3 Quality Assurance

Validation and auditing represent two independent phases of quality assurance for material deposited within the Archive. Ingest validation is performed to insure that content submitted for deposit is syntactically correct with respect to the published standards of the Archive. Additionally, validation will attempt to determine the correctness of submitted metadata (e.g., does the ISSN match the journal title and do data files actually conform to their specified formats) and the internal consistency of individual content objects (e.g., are all article bibliographic references correctly associated with the citations in the body of the text). No SIP will be accepted into the Archive until it has successfully passed ingest validation. The responsibility for correcting errors uncovered during ingest validation will rest with the submitting

publisher as will be specified in the Archive submission agreement. To lower operating costs and to facilitate the effective scaling of Archive operations, ingest validation will be automated to the fullest extent possible.

Since materials deposited within the Archive will generally be dark with respect to access for some initial period of time, it is important to allocate substantial effort to the validation of the quality of submitted material upon ingest. The difficulty, and thus cost, of the identification and correction of errors in archived journal content will only increase over time. We will develop tools to perform automated quality assurance (QA) testing at a syntactic level, and as far as practicable, on a semantic level. In addition to the internal use of these tools by the Archive, they will also be made available to content providers for client-side validation prior to submission. Thus, these systems will be implemented with regard to platform independence, and keeping in mind the wide range of technical resources available to potential content providers, ease of installation and use.

Due to the wide variety of publisher production workflows and content management systems, and the fact that item-level content is submitted to the Archive in a common schema, produced by transformation from its native form, it is important for ingest QA testing also to include semantic level validation. All online content providers rely to a greater or lesser extent upon the high degree of domain expertise of their users in detecting semantic errors. As material submitted to our Archive will remain dark for some initial period of time, relying solely upon this approach is not feasible. It is also not feasible to assume that the Archive staff itself can ever possess the same width and breadth of domain knowledge as is present in the scholarly community.

Our approach to this problem is to move semantic validation to the level of copy-editing. Within the SIP, content providers are asked to provide a rendered version of all item-level content in a standard page description format (e.g., PDF), derived from the provider's internal native form of the content which is presumed to be authoritative. After SIP ingest, a rendered version of the item-level content is derived from the Archive's common schema version of that content. Proofreading between these two versions will suffice to detect semantic errors. The scope and selection of material that is validated in this manner will be adjusted over time with regard to the detected error rate, perhaps on a publisher and title basis.

4.2.1.4 Descriptive Information

After SIP validity has been confirmed, issue and item-level descriptive metadata is extracted from the issue and item-level METS files of the SIP, and transmitted to the Data Management function for storage and later use in archive administration and resource discovery.

4.2.1.5 Transformation of SIP to AIP

Following validation and descriptive cataloging, the individual components of the SIP are transformed into the AIP format for deposit into the DRS in its capacity as the Archival Storage entity. For the most part, SIP components are deposited as is. The METS files are

rewritten to include additional internal archive-specific administrative metadata, and to change the references to content files from file references valid within the SIP file system hierarchy to DRS inter-object references.

Some of the technical metadata existing in the METS files may be duplicated in internal DRS storage structures in order to facilitate ongoing archive administration and preservation activities. Whenever feasible we will attempt to harvest technical metadata stored internal to submitted SIP components.

4.2.2 Data Management

The Data Management function is responsible for maintaining descriptive information about archival holdings and administrative data necessary to the internal management of the Archive. Issue and item-level descriptive metadata is received from the Ingest function. Issue-level metadata is stored in the existing HOLLIS ILS, which includes serial check-in and claiming mechanisms useful to detect and request submission of missing issues. Item-level metadata is stored in a new catalog, implemented either in the HOLLIS ILS or as a stand-alone XML database application.

4.2.2.1 Bibliographic control

E-journal content is modeled within the Archive at both an issue level and an item level. Issues are defined loosely as primarily publisher-specified aggregations of individual items, with some additional issue-level, and generally non-citable, content such as masthead, editorial board, cover image, etc. Items are defined as indivisible pieces of citable content such as articles, editorials, reviews, letters, errata, etc. For purposes of internal administration of the Archive as well as for end-user content discovery and delivery, bibliographic control of journal content is necessary at both issue and item levels.

This two level modeling scheme is explicitly issue-centric. While our publishing partners are interested in exploring the modality of issue-less publishing in the future, none of them have indicated that this will occur during the scope of this project. Thus, for the purposes of streamlining deployment we are maintaining our conceptual focus on issues while remaining cognizant of the fact that this is an area that will require additional work in the near future.

In keeping with the established policy of the Harvard University Library, no artificial distinction is drawn between analog and digital assets in library catalogs. As issue level information (e.g., title, ISSN, publisher, holdings by chronology and enumeration, etc.) is already being captured in the library's existing OPAC for print and online editions of serials, we will provide similar bibliographic information in the union catalog for archived e-journals. Discovery of archived content will use the standard search mechanisms provided by the Web-accessible OPAC. In addition, we will construct a new catalog for item-level bibliographic control specifically to capture and make searchable item-level information.

4.2.2.2 Naming

Naming is the process of assigning unique, persistent identifiers to resources. Uniqueness insures an unambiguous mapping between an identifier for a resource as described in a discovery service and the instantiation of that resource as delivered to the user. Persistence is required when identifiers for archived content are publicly visible and thus, susceptible to being captured and used in external systems outside the control of the Archive. A "bookmarked" name should always resolve to the correct named content object regardless of the passage of time or changes to the underlying architecture or implementation of the Archive. Within the Archive, naming needs to occur at the level of granularity of the discovery and delivery services, that is, at the issue and item levels.

The Digital Object Identifier (DOI) mechanism is the most widely used naming scheme for electronic journal articles.[14] However, DOIs resolve to resource instantiations defined by the registering body for those DOIs, in this case, by the publishers. Therefore, an article DOI will resolve to that article in the publisher's content delivery service. There is no widely implemented mechanism to interrupt the DOI resolution process and substitute resolution to a local "appropriate copy" such as the Archive. Thus, it is necessary for an Archive-specific identifier to be given to named issue and item-level content components.

The Harvard University Library operates its own Name Resolution Service (NRS), composed of an administrative registry of name-to-URL mappings and a resolution server compliant with established Internet Engineering Task Force (IETF) protocols for Uniform Resource Names (URNs). Syntactically, URNs are always generated with an internal namespace designation to avoid collisions between names assigned by different naming systems. The Harvard namespace, "urn-3," is registered with the Internet Assigned Numbers Authority (IANA). Although NRS names will be used in the Archive discovery and delivery services, Archive metadata for e-journal content objects will also include other public and private identifiers associated with those objects, including DOIs and publisher-specific internal identifiers.

4.2.3 Archival Storage Strategy

Our three publishing partners currently offer 1,137 electronic journals, annually comprising over 210,000 articles with a total size of approximately 400 gigabytes per year (assuming SGML/XML full text and TIFF images for articles, each with an accompanying PDF file). They project relatively modest growth in their electronic offerings over the next five years, with an anticipated increase in titles of three percent per year.

The current storage architecture underlying our operational Digital Repository Service (DRS) uses NFS-mounted, RAID-based devices as its primary online storage mechanism, with automatic replication over a dedicated T1 network to an off-site tape library which can be mounted as a file system for remote recovery access. The operational policy for the tape library enforces automatic periodic tape refreshment on a five year schedule. The

total growth capacity of the current implementation of this system is 50 terabytes. The recovery cost for storage under this system is $20/gigabyte/year.

The current practice of our publishing partners is to present journal content in the form of static text and visual images. As advanced dynamic media types, such as streaming audio and video, become more prevalent in electronic publishing, the per-issue size requirements of the Archive will increase commensurately.

The Archival Storage function of the Archive is provided by the extant DRS. The DRS batch loading process requires that each physical data stream be available as a separate file, along with an additional XML-encoded control file specifying loading and storage options. This proscribes the form of the AIP within the DRS as the complete set of SIP components with each issue and item-level METS metadata and content file deposited as individual digital objects. The Ingest function is responsible for transforming the SIP into the AIP prior to DRS deposit. After successful deposit of an AIP into the DRS, the Archive will generate and transmit an e-mail message of confirmation to the submitting content provider. The issuance of this confirmation constitutes the formal notice of the archive's assumption of archival responsibility for the deposited material.

Issue and item-level METS metadata files contain internal pointers to their component content files. In addition, the DRS has its own explicit mechanism to maintain typed relationships between individual digital objects stored within it. Thus, issue and item-level METS files will be stored within the DRS with an inter-level parent/child structural relationship. Similarly, METS metadata files and their component content files will be stored with an intra-level parent/child relationship.

Given the substantial number and size of e-journal components that will be deposited over time, we will allocate resources to evaluate the DRS's scaling properties and, if necessary, to design and implement appropriate enhancements.

4.2.4 Preservation Strategy

Because the purpose of this Archive is to preserve the significant intellectual content of journals — not the original form in which the content was authored or delivered — the Archive will most likely rely upon transformation to prevent obsolescence.[15] As defined by OAIS, *transformation* refers to "a [type of] digital migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content."[16]

Files that are proprietary and therefore not amenable to transformation will also be accepted into the Archive, provided they are not essential to the meaning of the journal article. The Archive will accept, store, locate, and deliver these files; the Designated Community would assume responsibility for transformation.

The Archive's preservation policies and management functions are format-specific, envisioned to meet the following objectives:

- to provide a range of preservation services according to what is viable for a given format at a given time[17]
- to monitor and document levels of technology support for file formats in a file format registry[18] that would:
- minimize the amount of technical metadata collected for each object
- promote collaboration among the Archive, industry, and standards bodies with domain expertise to define the "trigger" events to initiate transformation
- to minimize costs with batch processing operations for file validation, monitoring, and transformation
- to promote best practices for authors to create and submit journal articles to publishers

4.2.4.1 Preservation Planning

The central premise of the Archive's preservation policy is that viable preservation services vary according to the shifting, contemporaneous level of support that data formats enjoy with regard to standards and applications. Recognizing that forward data migration of archived e-journal content will not always be lossless and in some cases not even possible, our policy is being modeled on the assumption that the Archive will offer multiple levels of preservation service.

4.2.4.2 Levels of Preservation Service

Preservation planning is an area of active development. Although much more analysis is needed, some key distinctions have emerged in considering the technical and operational implications of assuming responsibilities to monitor obsolescence and to migrate data. Our expectation is that the Archive will offer multiple levels of service in which the highest level ("Level One") represents the Archive's commitment to monitor formats and associated technologies, to develop and execute migration strategies that attempt to preserve all of the format's native functions and semantic integrity, and to disseminate files (e-journal components) in formats that can be rendered by contemporary applications.

The Archive will provide the highest level of preservation service for a limited set of preferred normative formats as discussed previously. Objects submitted in non-normative formats are expected to fall into two categories: those that can be transformed upon ingest into an analogous normative format and those that cannot (e.g., files with encryption or proprietary compression). Objects in this latter category will receive fewer services. At a minimum, objects will receive "bitstore service" in which they are refreshed and can be disseminated from the Archive to members of the designated community or to "digital archaeologists" committed to investing the resources needed to re-render the objects with contemporary applications.

Challenges remain to define the terms and conditions in which objects will receive middle levels of preservation service — those that offer more than bitstore, but less than Level One — such as "lossy migration." The Archive's preservation policy will include statements that address how objects are classified upon ingest to receive specified levels of preservation service and what circumstances could lead to the service level being promoted or demoted over time.

4.2.4.3 Policy Implications

The implication of instituting a preservation policy with multiple service levels is that all objects associated with an e-journal item can be deposited to, and accepted by, the Archive, but those integral to the item's semantic meaning should when at all possible be deposited in normalized, repository-compliant formats for full preservation service. Our publishing partners support this concept and are eager to use this archiving policy as an incentive to motivate authors to submit content in fewer, standardized formats.

Within the Archive, a preservation manager will be responsible for monitoring data format-specific technological trends, as well as the needs and capabilities of designated user communities. To facilitate reliable monitoring and migration planning we will develop a comprehensive data format registry, an authoritative repository of format metadata — or in OAIS terms, representation information — including an authoritative specification, the organizational entity responsible for format maintenance, a list of key applications capable of reading and writing the format, and the technical attributes that represent the functional integrity of the format. These latter properties are of particular importance in modeling format migrations as their values can be used to distinguish lossless from lossy transformation. The assignment of high-level status to a particular format, for example, is due in part to the comprehensiveness of its representation information. We have developed preservation metadata requirements for XML, raster still image, and audio formats in the planning phase of this project; requirements remain to be defined for vector still image, page description, and other formats expected to be submitted to the archive.

Reporting functions will be developed to enable the preservation manager to track periodically the numbers of formats and format types, as well as the relationships among objects stored in multiple versions in the Archive. Procedures to identify potential technological obsolescence of selected formats and to present the costs and benefits of various migration options, when feasible, must also be developed to ensure that forward migration is always appropriately scheduled and performed in the most cost-effective manner.

We will also investigate the costs and benefits of preserving all versions of files (following each transformation) versus maintaining only the current version along with the technical metadata necessary to facilitate reverse engineering or, at the very least, a trail of useful provenance.

4.2.5 Access

The Access function encompasses both e-journal resource discovery and delivery. Discovery takes place at the title and issue level through the existing Web-accessible HOLLIS OPAC. Actionable issue-level links provide users with a table-of-contents-like view of all individual issue items. Actionable item-level links from the table-of-contents-like issue display and from search result records in the item-level catalog provide users with access to individual issue items. XML-encoded full-text content items are dynamically transformed into HTML using XSLT. Other item content media types (e.g., streaming audio or video) are delivered via the appropriate DRS delivery applications.

Rather than browsing or searching these Archive catalogs for relevant content, a user may possess *a priori* appropriate citation information for the desired content, such as author and title, chronology and enumeration, or a DOI uniquely identifying an item. A valid access request to the Archive can be made by using this citation information in a properly formed OpenURL which encodes the citation data into an actionable URL.[19] The Archive will implement an OpenURL service to accept such requests and respond with the appropriately displayed item.

Access authorization is a binary function. During the initial dark period following submission, e-journal content is available only to Archive staff for internal administrative and maintenance purposes, and to auditors as described in the Administration function. Once content is lit up in consequence of an appropriate trigger event, that content is available to the general public. Authorization is enforced only at the point at which the delivery of actual issue or item-level content is requested; cataloging metadata for all content is always available to the public.

Delivery of raw e-journal issue data (i.e., issue content and metadata as preserved in the Archive) can be requested and returned as an OAIS Dissemination Information Package (DIP). At least one form of the DIP should be equivalent to a Submission Information Package (SIP) so that a DIP delivered from a compliant archive can be ingested directly by another archive. The Archive will consider support for additional standard DIP formats as they emerge in the future. For the wholesale batch dissemination of archived content, that content will be transformed from its internal Archival Information Package (AIP) form to a DIP. The handling of and response to requests for a DIP will be an off-line, asynchronous operation. It is highly desirable that the archiving community cooperates in the development of standard definitions of the DIP to facilitate the transformation of archival materials between participating institutions.

4.2.6 Administration

The Administration function is responsible for the routine operation of the Archive. A good deal of this work is manual, not automated, including the negotiation of submission agreements with content providers, supervision of the Archive staff, and the maintenance and enhancement of the Archive's technical environment and infrastructure. Nonetheless, these manual procedures will be supported by systems providing administrative database

services, online registration of content provider profile information, and hardware and software monitoring tools.

The major automated task of this function is the performance of required format migrations as instigated by the Preservation Planning function. The Archive's adherence to the internal use of a limited set of normative Level One data formats constrains what would otherwise be a potentially intractable undertaking into a feasible task. We will investigate the use of commercial tools to transform non-normative formats into normative data formats eligible to receive Level One preservation services. Additional tools may be identified to perform other data management functions, such as validation and metadata extraction, that would assist preservation monitoring when transformation is not feasible. In addition to the transformation and validation of the e-journal content objects, such migrations may also necessitate enhancements to delivery systems. As the set of normative formats grows or is culled over time, ingest procedures will also require concomitant modifications.

Although extensive content quality assurance occurs upon Ingest, subsequent periodic auditing of archived material will also be carried out to validate the lossless nature of migration transformations as well as the general stability of the Archive's storage environment. This auditing will occur under the purview of the Administration function by domain experts drawn from publishers, scholarly societies, and librarians. Statistical sampling of the Archive's holdings categorized by publisher, subject area, and title will be employed in the selection of material to ensure appropriate representative coverage in the auditing process.

4.3 Schedule

A fair number of issues remain to be resolved. In order to fully test the model for the Archive and to get a better understanding of the real costs involved in operating the Archive, Harvard believes it is necessary to build and run the Archive long enough to gather experience. The work schedule, taking this into account, is composed of the following four main functional phases:

- A one year development period primarily concerned with building additional needed pieces of infrastructure, designing the common archival item-level schema, finalizing the details of the SIP, and allowing sufficient time for our publishing partners to develop appropriate export mechanisms for Archive submission.

- A six month ingest test period using a controlled submission of a limited number of titles from all three publishing partners to test and validate the stability and appropriateness of the Archive's technical systems and workflow processes.

- A one year production ramp-up period during which the Archive will steadily increase submission volume to include the complete list of titles available from all

three publishing partners. This phase will evaluate the scaling properties of the Archive's technical design and operational plan.

- A one-and-a-half year full production period to confirm the operational stability of the Archive.

5 Roles and Responsibilities

5.1 Internal Roles and Responsibilities

5.1.1 Technical Development

The primary responsibilities of the Archive staff with regard to technical issues are the initial development, ongoing maintenance, and future enhancement of Archive systems. Additionally, the expertise of the staff will be helpful in monitoring technical innovation and obsolescence with regard to normative data formats and potential preservation migration transformations. This activity would be performed in cooperation with the Archive preservation manager and collection curators. All technical work will be performed using accepted industry standards and processes for technical management, design, implementation and testing methodologies, configuration management, and documentation.

The Archive architecture relies heavily on efficiencies achieved by compliance to common standards, e.g., SIP structure and the archival schema. Widespread compliance is best achieved by insuring that these standards are developed in an open collaborative process. Archive staff will be responsible for the coordination of these processes and the resultant timely publication of appropriate specifications.

As Archive submission is opened to publishers beyond our initial partners, we anticipate working with institutions with widely varying technical resources and competencies. Potential difficulties in this regard can be mitigated by the distribution of appropriate utilities and tool sets to facilitate publisher activities. All Archive development will be evaluated in terms of the applicability of new system components for such distribution. If found to be relevant, development will proceed with due consideration towards platform independence of the system implementation. Archive staff will also be available for limited technical consultation regarding publisher development and operational procedures.

5.1.2 Archive Content Development

Harvard's initial approach to archiving has been "publisher based," that is, oriented towards archiving all of the e-journal output of a publisher. This approach was chosen for two reasons. First, it simplifies the task of creating a large base of archival activity to test systems and operations, and to provide enough scale on which to base long-term cost projections. Second, we believed that the marginal cost of archiving another title from a publisher with whom the issues of interoperation have already been worked out would be

less than taking a new title from a new publisher. The cost of archiving a title could thus be lowered.

Our plan is to work with our three publisher partners, pending final agreement, to build and test interoperation between the Archive and the publishers' systems, then begin to increase the number of titles archived from each. One of the great uncertainties of archiving is the amount of labor required for ingesting new content. The Archive will very likely be limited in staffing in the initial phase of its operation. We plan to increase Archive coverage until we are ingesting all of the content from our original partners, or until the available staff cannot deal with additional input. At the point where all available titles from the original publishing partners have been deposited successfully and it is determined that we have not yet reached our capacity to ingest content, it will be appropriate to evaluate the Archive's procedures and functions and determine growth options and extended partnerships.

5.1.3 Curatorial Responsibilities

In traditional preservation, curators are responsible for ensuring that collections remain usable. While they may store collections in centralized, environmentally-controlled storage facilities, or partner with conservators and preservation technologists to repair or copy materials, curators are ultimately fully responsible to account for the extent, condition, and usability of their collections.

The inherent fragility and complexity of digital collections require a shift in preservation responsibilities. To ensure that items do not become obsolete, curators and other owners need to delegate preservation responsibilities to technical staff with the expertise and the tools. In this new model, preservation technologists assume perpetual rather than temporary custody of the physical objects in their care. They must monitor the items as well as the environment. Because obsolescence is inevitable for all digital formats, they must be able to develop and present migration strategies to the curators and owners, then implement the strategy the owner prefers. In traditional preservation, curators have ongoing custodial responsibility (whether passive or active) and preservation technologists intervene infrequently. In the preservation model for digital archiving, repository managers assume the ongoing custodial responsibility and curators are consulted as necessary to make decisions about migration and other issues.

5.2 External

As we conceive this Archive, it cannot and does not stand in isolation. The Archive itself has a variety of partners and stakeholders. Additionally, the Archive must have a relation with the broader community.

5.2.1 Stakeholders

Discussion among Harvard and its publishing partners has centered on who "owns" the Archive and who governs it. While the Archive is intended to be maintained and

administered by Harvard and built on Harvard's existing digital library infrastructure, the publishing partners have suggested that a broader group with a vested interest should be involved. Who are the stakeholders and what is their role in helping the Archive do its job? This community could comprise authors, scholarly societies, publishers, and institutional subscribers as representatives of researchers. These delegated stakeholder groups should have a role in reviewing the policies and practices of the Archive as a mechanism for vetting the Archive and establishing a level of trust; however, some publishing partners have suggested that actual governance of the Archive is tied to the brightness of the Archive and the additional services that might be offered by the Archive. The brighter the Archive, the more governance a publisher should have; the more services offered, the more governance a publisher should have. Harvard, however, maintains that final governing authority is intrinsically tied to the ability to use its existing infrastructure as a starting point for the Archive, while a variety of policies and procedures related to the development, administration, ongoing maintenance, and financing of the Archive should be developed in consultation with and open for review and comment by representatives of this stakeholder community.

5.2.2 The Archival Community

In addition to the stakeholder community with its representative input, we have elsewhere in this paper discussed the necessity for an external auditing service. Such a service might be part of a broader confederation of archival organizations and stakeholders. Such a confederacy might be charged with establishing registries for content and format types; certifying policies, practices and procedures; and supporting the ongoing development of digital archiving.

5.2.3 Sharable Infrastructure

During the development of the Archive several pieces of sharable infrastructure will be produced:

- Format registry to provide a centralized store for relevant information about data formats supported by the Archive.
- SIP/DIP specifications. Community-wide agreement on these specifications will allow the free interchange of archived materials between archiving institutions and projects.
- Issue-level content schema to capture issue-level information such as masthead, editorial board, etc.
- Canonical item-level schema designed to accommodate the Archive's need for homogeneous content and allow a clean transformation path from publishers' native content formats.
- METS Java toolkit for API-level support for the procedural construction, validation, and serialization/deserialization of syntactically valid METS files.

- SIP Quality Control tool.

- XLST stylesheets for issue and item-level display, based on the specifications for the issue and item-level XML schema.

6. Postscript: 2003

Since the conclusion of our Mellon-funded planning project, additional work on e-journal archiving has continued. Through our investigation of potential e-journal XML schemas we learned that the National Library of Medicine (NLM) was in the process of revising the document type definition (DTD) used for PubMedCentral (PMC). Since PMC receives content from various publishers, their DTD has to support heterogeneity of document structure, although all the content is within the biomedical discipline. NLM was receptive to our suggestion that the DTD be expanded to support content across academic disciplines. Working with two leading XML consulting firms, Inera Inc. and Mulberry Technologies, Inc., with additional design input from Harvard, NLM has recently released an archive and interchange DTD suite (http://dtd.nlm.nih.gov/) to provide a common format in which publishers and archives can exchange e-journal content.

The main benefits of the DTD suite are:

- It was not created by, not reflects the bias of, any specific publisher, society, typesetter, or aggregator

- The document analysis covered a wide range of academic disciplines to insure that the DTD is not biased towards any specific intellectual domain

- It is based on public standards, such as the CALS and XHTML table models, MathML, and Unicode

- It is modular and can be modified easily to meet specific needs without undermining either the code structure of the DTD or the interchange of files created according to the DTD

- It was designed so that publishers can easily transform existing content to be compliant with the DTD

Although just recently released, the DTD suite is undergoing serious evaluation and prototypical use by many content providers and suppliers. If the DTD suite fulfills its promise, it will provide for the foreseeable future a common scholarly publishing DTD for purposes of article interchange and archiving. Within archival repositories, significant economies of scale can be achieved only through large-scale automation, which in turn requires maximum homogeneity of content. The NLM DTD suite fostered by the work of the Harvard University Library should help facilitate the creation and operation of sustainable archives for e-journals and thus help to promote their use within scholarly pedagogy and discourse.

7. Endnotes

[1] Dan Greenstein and Deanna Marcum, "Minimum Criteria for an Archival Repository of Digital Scholarly Journals," Version 1.2 (Washington, DC: Digital Library Federation, 15 May 2000). Online at http://www.diglib.org/preserve/criteria.htm. Also available in this publication.

[2] Anne J. Gilliland-Swetland, *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment*, Publication 89 (Washington, DC: Council on Library and Information Resources, February 2000). Online at http://www.clir.org/pubs/reports/pub89/contents.html.

[3] It is possible for typographical manifestation to impart significant semantic value, as in the case of poetry and other forms of creative expression. In such cases where the manifestation forms an intrinsic part of the intellectual content, we will explore mechanisms to identify, capture, and preserve this in the Archive.

[4] For information about GenBank, see http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html.

[5] In rare cases, an article included in the print version of a journal issue is not available in electronic format. The fact that such as article is not available should be noted.

[6] Harvard University Library, *Library Digital Initiative Home Page* (Last modified January 2003). Online at http://hul.harvard.edu/ldi/.

[7] Internet Engineering Task Force (IETF), *Uniform Resource Names (urn) Charter*, 55th IETF Working Group Meeting, Atlanta, Georgia (Last modified 31 July 2001). Online at http://www.ietf.org/proceedings/02nov/126.htm.

[8] R. Daniel, *A Trivial Convention for using HTTP in URN Resolution*, RFC 2169 (June 1997). Online at http://www.ietf.org/rfc/rfc2169.txt.

[9] Consultative Committee for Space Data Systems, *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1 Blue Book (Washington DC: National Aeronautics and Space Administration, January 2002). Online at http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf.

[10] Network Development and MARC Standards Office, Library of Congress, *Metadata Encoding & Transmission Standard (METS)*. Online at http://www.loc.gov/standards/mets/.

[11] Harvard University Library, *Submission Information Package (SIP) Specification*, Version 1.0 DRAFT (19 December 2001). Online at http://www.diglib.org/preserve/harvardsip10.pdf.

[12] Inera, Inc., *E-Journal Archive DTD Feasibility Study* (5 December 2001). Online at http://www.diglib.org/preserve/hadtdfs.pdf.

[13] John Mark Ockerbloom, "Archiving and Preserving PDF Files," *RLG DigiNews* 15.1 (15 February 2001). Online at http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2.

[14] International DOI Foundation, *The DOI Handbook*, Version 1.0.0 (February 2001). Online at http://www.doi.org/handbook_2000/010222DOI-Handbook-V100.pdf.

[15] Such transformations can be either anticipatory, with the repository maintaining the transformed version of the object, or "just-in-time," with transformation happening when a user requests an object.

[16] *Reference Model for an Open Archival Information System* (OAIS): 5-5. URL at note 9 above.

[17] Preservation levels may be negotiated according to changes in the level of support for a given format by standards, industry applications, or applications within the Archive.

[18] Harvard and MIT have begun discussing a registry framework that will potentially be shared by both electronic journal archives.

[19] Herbert Van de Sompel, Patrick Hochstenbach, Oren Beit-Arie, "OpenURL Syntax Description," version OpenURL/0.1f (16 May 2000). Online at http://www.sfxit.com/openurl/openurl.html.

8. Appendix A: Project Staff

8.1 Project Steering Committee

This group was composed of senior curators, preservation experts, and library systems staff to address functional and organizational issues. Members of the Committee are:

Ivy Anderson, Coordinator for Digital Acquisitions

Marianne Burke, Assistant Director for Resource Management, Countway Library of Medicine (January 2001-September 2001)

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

Diane Garner, Librarian for the Social Sciences, Harvard College Library

Marilyn Geller, Project Manager (July 2001-March 2002)

Jeffrey Horrell, Associate Librarian of Harvard College for Collections

John Howard, Associate Director for Technology Development & Services, Countway Library of Medicine (September 2001-March 2002)

Y. Kathy Kwan, Project Manager (January 2001-June 2001)

Jan Merrill-Oldham, Malloy-Rabinowitz Preservation Librarian

Constance Rinaldo, Librarian, Ernst Mayr Library of the Museum of Comparative Zoology

Lynne Schmelz, Librarian for the Sciences, Harvard College Library

MacKenzie Smith, Digital Library Projects Manager (January 2001-December 2001)

8.2 Project Technical Team

An internal team composed of staff with significant experience in digital library development investigated technical issues and systems requirements. Members of the Team include:

Stephen Abrams, Digital Library Software Engineer

Stephen Chapman, Preservation Librarian for Digital Projects

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

Marilyn Geller, Project Manager (July 2001-March 2002)

Y. Kathy Kwan, Project Manager (January 2001-June 2001)

MacKenzie Smith, Digital Library Projects Manager (January 2001-December 2001)

Robin Wendler, Metadata Analyst

9. Appendix B: Titles Included in E-journal Component Survey

Title	Publisher	Print	Electronic
American Journal of Human Genetics	University of Chicago Press	Yes	Yes
Acta Zoologica	Blackwell	Yes	Yes
American Journal of Physical Anthropology	Wiley	Yes	Yes
Astrophysical Journal	University of Chicago Press	Yes	Yes
Current Issues in Education	Arizona State University and the College of Education	No	Yes
Electronic Journal of Combinatorics	Neil J. Calkin and Herbert S. Wilf (in association with American Mathematical Society)	No	Yes
ENDS Environment Daily	Environmental Data Services Ltd	No	Yes
European Journal of Organic Chemistry	Wiley	Yes	Yes
First Break	Blackwell	Yes	Yes

Fish & Shellfish Immunology	Academic Press	Yes	Yes
Journal of Internal Medicine	Blackwell	Yes	Yes
Journal of Political Economy	University of Chicago Press	Yes	Yes
Journal of Seventeenth Century Music	Society of Seventeenth Century Music	No	Yes
Journal of the History of Behavioral Sciences	Wiley	Yes	Yes
Medieval Review	The Medieval Institute, College of Arts and Sciences, Western Michigan University	No	Yes
Nature	Nature Publishing Group	Yes	Yes
Nursing	Blackwell	Yes	Yes
Philosophy	Blackwell	Yes	Yes
Politics Research Group Working Papers	JFK School of Government, Politics Research Group	No	Yes
Representation Theory	American Mathematical Society	No	Yes
Science	AAAS	Yes	Yes

10. Appendix C: Electronic Journal Archives Survey

Thank you for agreeing to participate in this survey for the Electronic Journal Archiving Project at Harvard University.

The Harvard University Library and three major publishers of scholarly journals — Blackwell Publishing, John Wiley & Sons, Inc., and the University of Chicago Press — have agreed to work together on a plan to develop an experimental archive for electronic journals. The preservation and archiving of electronic journals — which are increasingly digital only and for which, in many cases, no paper copies exist — present unique, long-term challenges to librarians, publishers, and ultimately, to the scholars and researchers who will seek access to them over time.

The new joint venture is sponsored by the Andrew W. Mellon Foundation which recently awarded a grant to the Harvard University Library specifically for the planning of an electronic journal archive. The year-long planning effort will explore the issues related to electronic journal archiving and develop a plan for a repository at Harvard for electronic journal publications. The expected outcome is a proposal for an archive for these journals.

We are currently exploring which components of e-journals can and/or should be archived. It is clear that all articles will be archived as well as reports, columns, editorials, communications, abstracts, errata, and correspondence. It is less clear which other components should be preserved and how non-traditional contents, such as links, data sets, and data simulations, should be handled. How much of the look and feel of an electronic journal should be preserved? Assume that not everything can be preserved because the costs will be prohibitive.

Another significant issue to be determined is at what point items in the archive may be accessed. For example, should the archive be "dark" and accessible only under emergency conditions (such as a publisher going out of business) or should the archive be "light" and effectively serve as an alternative to a publisher site for everyday access to journals? Most likely, the scenario will be something in-between.

For this survey, please consider that the need for a particular journal component will be no sooner than ten years in the future. Also assume that not all journal features can be preserved.

SURVEY QUESTIONS

Please rate the importance of the following components of an electronic journal. (The concept *future* is defined as ten or more years into the future.) Circle 1, 2 or 3 with the meanings:

1 **no future use is likely**

2 **limited future use is likely**

3 **important to maintain future access to this component**

Journal Content

1 2 3 Cover image for issue
1 2 3 Table of contents
1 2 3 Volume/issue number linked to content
1 2 3 References to outside information (e.g. portals, author-developed data that is stored outside the journal site, bibliographies developed by the journal, etc.)
1 2 3 Threaded discussion
1 2 3 Index to volume
1 2 3 Subject
1 2 3 Author
1 2 3 Advertisements
1 2 3 Announcements (e.g., events)
1 2 3 Editorial board
1 2 3 Editorial policy
1 2 3 Membership list (e.g., for societies)
1 2 3 Reviewer list
1 2 3 Copyright
1 2 3 Licensing information
1 2 3 Guidelines for authors (e.g., manuscript preparation and submission)
1 2 3 Business information
1 2 3 Advertising guidelines
1 2 3 Subscription information
1 2 3 Customer Service
1 2 3 Contact information
1 2 3 Career/job information

Journal Functionality

Browsing

1 2 3 Chronologically
1 2 3 By subject
1 2 3 Links within the journal, volume, issue (e.g., article to article links and links to supplementary information within the journal)

Searching

1 2 3 Author
1 2 3 Title
1 2 3 Keyword
1 2 3 Limit by date
1 2 3 Help
1 2 3 View thumbnail of image
1 2 3 View full-size image

11. Appendix D: Archive Workflow