Report on a Mellon-Funded Planning Project for Archiving Scholarly Journals

John Mark Ockerbloom
University of Pennsylvania Library
16 September 2002

Table of Contents

Abstract

We report on the Penn Library's Mellon-funded project to plan for electronic journal archiving. Our project focused on working with two university-affiliated publishers, Oxford and Cambridge. We discussed rights and responsibilities, built prototypes for automatic importation of presentation files and metadata, studied issues of migration, particularly of PDF files, and considered archival economics and tradeoffs between funding by publishers and access by libraries. We recommend a broadly supported archival system and outline two architectures for such a system: a centrally managed archive, supported by libraries and giving libraries access to archived material (extending the JSTOR model),[1] and a widely distributed archival network of library-based lightweight repositories (extending the LOCKSS model).[2] For both models, we discuss how service providers in the architecture can carry out important archival support tasks while easing the workload on central servers or on library repositories. We see Penn's immediate role in the archival community not as an integrated archive, but as one of the service providers for key functions including ingestion, migration, and registry services.

Introduction: Why are we doing this?

Digital and network technologies are rapidly changing the nature of scholarly communication. Scholars increasingly use the electronic medium to publish their research results. In this medium, new knowledge can be disseminated extremely quickly by anyone with access to the Internet, and accessed at any time from anywhere with a network connection. Costs to reproduce and disseminate information — though not necessarily the costs to produce, review, and edit it — can be greatly reduced online, thus enabling organizations that could not afford to publish scholarly information to do so. The nature of the electronic medium also supports new modes of publishing. Freed from the constraints of print, scientists can publish long, complex data sets along with their articles. Sound, video, and interactive programs can be published along with text and still images. Search and linking facilities can be added to electronic versions of journals and monographs making it substantially easier for researchers to locate and use relevant research. The interconnectedness of the Internet also allows the nature of journals to change from a set of issues with a static set of articles, to more dynamic collections and databases of published research.

Consumers as well as producers of scholarly content have embraced the electronic medium. A 2002 survey of Penn library users revealed that users overall consider the library's provision of electronic information needed for work more important than its provision of print information needed for work. While some communities, particularly those in the arts and humanities, ranked providing print information as higher priority, others, particularly users of Penn's biomedical, engineering, and business libraries, ranked electronic resources as significantly more important.[3] Electronic journals and databases that describe journal articles (often with links to electronic full text) rank among the most frequently accessed resources in Penn's libraries. In June 2002, the Penn libraries carried over 6,500 full-text online journals, over 5,600 of which were paid

subscriptions. The most popular of these electronic journals was accessed over 15,000 times per year by Penn researchers, and several others were accessed over 5,000 times per year as well.[4]

While electronic journals offer new benefits, they also bring new risks. Researchers depend not only on current journals, but on the centuries-old back files of journals that they can use for scholarship. However, electronic journals are not easily preserved. Unlike print journals they live on fragile media, rely on rapidly changing technology and computing environments for their presentation, and depend on the policies and fortunes of their copyright holders for their disposition. The new types of content supported by electronic journals, the advanced search and linking facilities, and the rapid access that electronic journal readers are accustomed to, further complicate the preservation problem.

Despite the complexities of the preservation problem, though, libraries cannot afford to ignore it. Electronic versions of scholarly journals are in many cases becoming the versions of record, with important scholarly communication not included in print publications. This includes not only multimedia content, but the interactive feedback that appears in journals intended for rapid dissemination of new results, such as BMJ, the online version of the *British Medical Journal*.

Understanding the progress of scholarship will require continued access to this record. At the same time, as other types of work of scholarly interest also move into the electronic realm — everything from email to researchers' notes to experimental prototypes to interactive Web sites — libraries and archives will have to find ways to preserve relevant materials in those areas as well. In some respects journals, which tend to have more predictable formats and publication cycles as well as more careful editing than other types of electronic works, are a useful starting point for understanding the general problem of electronic preservation.

We at the University of Pennsylvania Library, then, decided to start planning for electronic journal archiving with the aid of a Mellon Foundation grant. In this report, we describe our experiences in the planning project, and summarize our findings and recommendations for electronic journal archiving. Originally, we planned to set up a journal archive for particular publishers' journals. By the end of the grant period, we felt it was best not to commit to a journal archive, but we do see a place for us as a trusted service provider for certain specific archival functions in an archival network. Some of our experiences in the planning project may also help us in the design of institutional repositories which may eventually play important archival functions as well.

The report is structured as follows: After a brief summary of our past arrangements for journal preservation in the print-only era, we summarize the basic requirements for a reliable electronic journal archive. We consider alternative forms of archived material and for the organization of archival communities. We discuss appropriate rights and responsibilities of archival systems. We then report on a few key points concerning the life cycle of the archival process, focusing in particular on ingestion and adaptation to changing technologies. Finally, we conclude with a summary of our experience and

recommendations for the next stages in developing the electronic journal archiving community.

We will address these issues not only from a theoretical perspective but also by reporting our own experience in the planning stage. The planning process made it clear that the challenges for electronic journal archives involve technical, social, and economic issues throughout the archival process. We often have to consider tradeoffs between what researchers would desire in an ideal archive and what is economically feasible in a reliable, robust, and sustainable record of scholarly communication. Also, we have to consider how an archive interacts with other archives and with other stakeholders in the archival community: authors, readers, educational institutions, and publishers.

Journal preservation in the print era

When considering strategies for preserving electronic journals it is worth asking whether special archives are needed at all. After all, in the print era libraries have preserved many scholarly journals through their normal operations without having to set up special archival units or to take on formal responsibilities for preservation beyond their own local communities. Libraries simply retained and bound copies of the journals they originally received from the publishers and kept them on their shelves. While these bound volumes can deteriorate physically over time or be subject to loss or damage, bound volumes that are over one hundred years old and still perfectly usable are not uncommon in research libraries. No copyright restrictions prevented libraries from binding and retaining old issues; no technological changes made these print journals unreadable. Moreover, libraries could exploit redundancy to fill gaps in their own holdings. Since copies of scholarly journals were sent to multiple universities, many of which attempted to retain them, a library lacking an issue a scholar sought could simply borrow the issue or obtain a photocopy of a needed article from another institution that retained a usable copy. Similarly, when journals and other materials were copied to microform, these microforms also were distributed to many libraries. Through these means, libraries could build up a robust preservation mechanism where the only costs beyond initial acquisition of materials were keeping them on the shelves, maintaining interlibrary loan systems, and sometimes acquiring "migrated" forms such as microfilms.[5] Since the libraries owned the journals and the first-sale doctrine of copyright law let them dispose of these journals as they saw fit, their preservation was also largely unconstrained by the journals' copyright holders.

The same approach is not unthinkable for electronic journals, but at present few libraries retain electronic journals and provide them to their readers. As long as this is the case, an equivalent informal substitute for archiving will not suffice. Most libraries, after all, can provide access to electronic journals simply by pointing to copies on publisher or aggregator sites, without the expense involved in acquiring and storing them locally. Therefore, unless libraries see definite benefits in keeping local copies that outweigh the costs, redundant local storage networks are unlikely to arise on their own, especially if they lack the durability and freedom to copy of the print journal environment. Therefore,

libraries will either need to organize or support archival organizations, or develop viable local storage networks that can reliably and cost-effectively preserve electronic journals.

An analysis of the informal print journal "archiving" network needs to consider its limitations as well as its advantages. Access to back issues of print journals is much slower than access to electronic journals, especially as libraries increasingly move older journal volumes out of primary stacks and into high-density storage. When researchers seek journal issues their institutions don't have, they may have to wait weeks for interlibrary loan. Cumulative storage costs for a journal volume redundantly stored in dozens or hundreds of locations is also significant, taken as a whole, even if the costs per institution are low. Furthermore, there is no guarantee that issues will remain obtainable.

Consider the experience of JSTOR which has digitized back issues of important scholarly print journals in the arts and sciences, with back issues provided by publishers and participating libraries. With over 1,300 participants in June 2002, JSTOR still lacked complete runs of 33 of its 218 online journals, with approximately 300 missing volumes and about 70 missing individual issues. About half of the missing volumes, and most of the missing individual issues, were not from the start of a run: that is, they represent gaps in holdings made available for digitization after libraries had started subscribing to the journals. If an average volume has six issues, these figures suggest a loss of about one-and-one-half percent of back issues for journals of interest to the JSTOR consortium, counting from the earliest available volume of journals, or a little over three percent counting from the first published issue. Many of these missing issues may well be in libraries and simply have not been made available to JSTOR. However, such issues are also likely to be relatively difficult to obtain by a scholar who might need them. (We have not yet, however, tried obtaining "missing" issues ourselves via interlibrary loan.)[6]

Requirements for electronic journal archives

The alternative to having libraries depend on their own and their neighbors' stored copies of electronic journals is to have them use trusted digital journal repositories. Several types of organizations could run such repositories, including the publishers of the content, particular academic or national libraries that state a commitment to preserve content, dedicated archival organizations, scholarly societies, or organized consortia of any of the preceding organizations. Whatever the nature of a repository's maintainer, however, users will have to trust that the repository will preserve content and provide access to it when needed.

Several papers have been published outlining the requirements for a trusted digital repository. In 2000, a Digital Library Federation working group that included both librarians and nonprofit publishers released a set of minimum criteria for a journal archival repository.[7] Here is a condensed paraphrase of these criteria:

- A repository will be a trusted party that conforms to minimum requirements agreed to by both scholarly publishers and libraries.

- It will define its mission in regard to the needs of scholarly publishers and research libraries, and be explicit about what it is willing to archive and for whom.
- It will negotiate and accept appropriate deposits from scholarly publishers.
- It will have sufficient control of deposits to ensure their long-term preservation.
- It will follow documented policies and procedures to ensure preservation in all reasonable contingencies.
- It will make preserved information available to libraries under conditions negotiated with publishers.
- It will work as part of a network.

What does a repository need to do to be considered a "trusted party"? A more recent paper issued by the Research Libraries Group specified these key attributes of a trusted repository (rearranged from their original ordering):[8]

- Administrative responsibility
- Organization viability
- Financial sustainability
- Technological and procedural suitability
- System security
- Procedural accountability
- Compliance with the Reference Model for an Open Archival Information System (OAIS)

The first three of these attributes deal almost entirely with the persistence of the organization, rather than any technical criteria. To be trusted, the organization must commit to being responsible for the archived material and be able to sustain this commitment over the long term, including any necessary financial commitments. Implicitly, this commitment is being made to all of the people who "trust" it: in this case, the scholarly community at large, including the publishers of scholarly material.

Libraries contemplating creating and maintaining a trusted digital archive, then, must consider such a commitment carefully. For many academic libraries, including Penn's, it represents a subtle but significant augmentation of a library's mission. In general, the Penn library is charged with maintaining and providing access to information for the teaching and research needs of the Penn community. It receives the bulk of its funding from the various schools and departments of the university that make up this community.

If material that the library has acquired no longer is of interest to the community, or is too expensive to maintain compared to other priorities of the university, the library can decide to dispose of it. Indeed, to properly carry out its mission, the library *should* dispose of it if this is necessary to sustain the library's more critical services, and if it has not made promises to others to retain the information.

This does not mean the library should not care about long-term preservation or should not participate in communal archiving efforts involving multiple institutions, but the library does need to carefully consider any commitment made to constituents outside the university to make sure it complements university interests. Because scholarly journals may have a usage life of centuries or more, the library needs to be sure it can sustain this commitment over such a timespan or can transfer its responsibilities to another trusted party if it cannot. It must also be sure it can continue to get funds to support this commitment. If its funds come from the university, the university as well as the library must be committed to the archiving project at the highest levels. If the funds come from other sources, such as archive depositors or users, the business model for the archive must show that these sources are sufficient to support the archive's commitments.

The more technically-oriented attributes of the RLG list emphasize the importance of prudently chosen, well-documented, secure, and transparent processes for archiving materials. The requirement to conform to OAIS provides a partial functional specification of such processes, using an established and tested ISO standard that details the archiving process and provides an information model for the data maintained as part of this process.[9]

One of the most important aspects of the OAIS model is its modularity. Archived materials in OAIS, and their metadata, come in Archival Information Packages (AIPs) and are likewise acquired and disseminated in similar packages. The functions of the archive are also divided into modules operating on the AIPs in various ways. This model allows us to look at an archival scheme not just as a large, monolithic unit, but as a system whose functions and responsibilities can be disseminated and replicated through a community of archival stakeholders. Content can be replicated reliably at multiple sites by replicating its AIPs, which contain all of the information needed to access and use the content. Archival functions such as ingestion and migration, because of their modularity, can also be delegated to multiple service providers (which may be separate from the rest of the archive) with appropriate expertise, certification, and trust. The model, in short, can apply to an archival system as a whole, not just to a single repository. In some cases, as we saw with the redundant informal network of print journal "archiving", the entire system can have properties that make it trustworthy, even if no individual repository in the system is particularly trustworthy.

What to archive?

Archives need to decide which of the thousands of electronic journals they will preserve. Two methods of narrowing down the field occurred to us, each with its own advantages:

- We could limit the number of publishers we would work with. This has several appeals. It limits the number of publisher agreements that need to be made, each of which may require substantial negotiation. It allows us to work with and support publishers with whom we already have close relationships and wish to see represented in the archive. It is likely to reduce the number of input formats to the archive, and the diversity of procedures required to ingest journal content, thus lowering overall costs of ingesting material. Furthermore, in a networked environment, with multiple, independently run archives, responsibilities can be more cleanly partitioned by publisher than by less definite criteria like subjects.

- We could limit the topical scope of the journals we would work with. This also has its advantages. It allows us to focus on disciplines that are of special interest to the university, and ignore those that are of little or no interest. It makes it easier for us to find experts in the disciplines that can advise us on which journals are most important to preserve, and on how they should be preserved. If the archive is sufficiently comprehensive, the archive can also serve as a self-contained research resource in its own right for its specialized disciplines. Such a resource can be "marketed" as a resource for the community engaged in that discipline, which may then fund the archive through subscriptions or grants.

We decided initially to only work with a few publishers. In another Mellon-funded project, we are working with Oxford University Press and Cambridge University Press to put history books online. We decided it would be fruitful to work with these publishers for journal archiving as well. We already had relationships with them, they published many high quality journals that Penn already subscribed to, and we wanted to support the role of university-affiliated publishers as important participants in the scholarly communication process.

We also needed to decide what constituted preservable content in the journals we handled. In online journals, which may have external hyperlinks, auxiliary searching and browsing scripts, dynamically updated and targeted advertisements (which may be out of the control of the actual journal publisher), comment boards, and ephemeral description files, the distinction between journal content that should be preserved and external content can blur. Since the purpose of the archive is to preserve the record of scholarship, though, the archive should at minimum preserve all refereed editorial matter published by the journals we archive. This includes not only research articles, but also reviews, letters to the editor, and other edited content. We considered it best not to guarantee archiving or preservation of advertising matter, ephemeral and informal indexes (such as a page with links to current and past journal articles), search engines, other server programs, or auxiliary materials not clearly part of the journal itself. Hence, an edited Letters column, published as part of an issue, would be included in the archive, an auxiliary, unmoderated threaded Web-discussion board would generally not need to be included in the archive. Also, while advertising may have some interest to future historians, it is not itself part of the record of scholarship, so it is not as important to an archive of scholarly journals as it might be to an archive of popular or trade journals. Technical and legal issues around downloading and archiving advertisements make them problematic to commit to

archiving them, except in cases where advertisements and scholarly content are integrated in the same file.

Exceptions may need to be made to the rules above for certain journals. For example, unrefereed submissions are an important part of some journals that focus on the latest progress reports of research. For example, the BMJ, the online version of the *British Medical Journal*, allows appropriately credentialed readers to make "rapid responses" to articles, which are usually posted within twenty-four hours of receipt. While these responses are moderated for relevance, they are not refereed, and editing is minimal. But they represent an important part of the scholarly communication provided by this online journal. An archive of this journal, then, should attempt to preserve these responses, even if archiving an unpredictably-growing set of responses to an article is more difficult to manage than a static set of articles originally published in a journal issue.

Archives also need to decide on the form in which they will preserve journal content. Will they preserve content as readers see it ("presentation" forms), or as their publishers prepare and structure it ("source" forms), or both? Each has its advantages and disadvantages, but a comprehensive archival system should ideally preserve both.

Presentation forms. In the case of print journals, libraries save the journals in the form in which they are presented: bound, printed pages. Electronic journal archives can also save such "presentation" forms. For many electronic journals these presentation forms consist of PDF or HTML and image files that are delivered to readers when they access journals over the Web. These files faithfully reproduce the appearance of journal articles and their text. They are also in many cases automatically harvestable from publisher sites — unlike source files which may not come in as consistent formats as presentation forms — and are therefore relatively inexpensive to ingest.

However, presentation files usually lack the structural markup available in the publisher's source forms. They also do not include any server-side scripts or other functionality that would be provided on the original publisher's Web site. Hence, they might not support functions that readers of electronic journals increasingly expect, such as automatic bibliographic linking, cross-article text searching, and data analysis. Migration to other formats may also be problematic, at least for anything beyond image capture and simple text streams, particularly if a nonstandard format is used. However, we believe it is important for archives to include presentation forms of journals because they show what readers actually saw when they read the journals.

Source forms. Electronic journals typically also produce "source" files which are the edited documents on which the presentation forms are based. For many electronic journals these are SGML files — or more recently XML files — following a structure known as a document type definition (DTD) specified by the publisher. The structure and markup of source files often provides information that cannot practically be extracted from presentation files, such as the structure of equations and formulas, bibliographic records, and tabular data. If the source files use a well-documented DTD for this information, programs can easily analyze journal articles to support value-added services.

On the other hand, archiving only the "source" forms means that viewers of archived journals might not see exactly what readers of the original journals saw. It may be difficult to verify whether the source materials can generate the same content and functionality of the original presentation forms, in part because source forms may not conform exactly to a standardized DTD, and also because publishers may make last-minute edits in presentation files that do not appear in the source files. There is currently no common standard used for various publisher DTDs, although many of them descend from a common ISO standard. A Harvard-sponsored study prepared by Inera recommends a common DTD that incorporates many of the elements of major publisher DTDs.[10] If such a standard were widely adopted it would substantially reduce the costs and uncertainties of maintaining source files.

Auxiliary forms. One of the appeals of electronic journals is that they can include digital content such as multimedia, programs, or machine-readable data sets that augment the text and illustrations of a traditional scholarly article. These supplements should be preserved in journal archives as well. However, because they often appear in complex formats that may have limited long-term support, an archive may not be able to guarantee that the supplementary content will continue to be usable by future researchers indefinitely. It may be useful, though, to support some basic supplementary data types, such as character and numeric data stored in relational tables. For other formats, if the supplementary material can be packaged as a bit-stream, archives should at least be able to preserve this bit-stream. They should also include some metadata concerning the format of the bit-stream so that new programs can interpret that format, if researchers are sufficiently motivated. A journal archive itself need not record detailed information about the format specification, so long as it can refer to a reliably preserved copy of the format specification stored elsewhere.

How should archives be organized?

Any archive that we construct will not stand on its own. Electronic journals will be archived at multiple sites, hence the DLF's requirement that journal archives operate in a network. In a well-designed archival network, institutions can share the burden of archiving journals, and libraries and researchers needing to use archival materials can locate and retrieve appropriate copies from a participant in the network. However an archival system is set up, though, it must be clear who takes responsibility for the material being archived and how this responsibility is enforced.

Self-archiving. Some authors and organizations may attempt to archive their own material. A publisher may declare that it will archive its own back issues indefinitely, for instance. Many authors have also provided preprints of their papers and journal articles on their Web sites for years. This informal practice is now encouraged by manifestos like the Budapest Open Access Initiative.[11] "Self-archiving" can also be done by authors' institutions. For instance, MIT's DSpace repository promises stable long-term storage for the scholarly work of its faculty.[12]

Self-archiving, however, will not suffice to preserve scholarly journals, let alone scholarly communication as a whole. It essentially relies on the self-interest of the original creators and publishers to keep the archive viable, but Web sites often disappear when an individual changes institutions or careers. Copies of papers on individual sites are often preprints, so they may lack important revisions and supplementary information that appeared in the journal version. Publisher-run archives may disappear without warning if they are no longer cost-effective to the publisher, or if the journal or the publisher fails or is acquired by another company. Contents of archives may be changed, corrupted, or withdrawn, either by accident or intentionally. As holders of copyrights to the content they archive, publishers and authors have exclusive rights to disseminate this content unless they grant rights to archives run by disinterested third parties. With exclusive rights comes near-exclusive responsibility for their material. Without very strong certification and backup strategies, self-archiving with this level of exclusivity is not likely to be widely trusted to preserve scholarly information for the time spans researchers and libraries have come to expect. This does not mean that self-archiving is useless, however. Self-archiving can be a backup to trusted archives run by other parties. Also, self-archiving, particularly the institutional self-archiving provided by systems like DSpace, may preserve information of scholarly interest that is not included in standard journal archives.

Integrated responsibility. Whether run by content creators and publishers or by third parties, archives have traditionally taken full responsibility for the content they archive. They are responsible for quality control, continued preservation, and providing appropriate access. If they themselves cannot maintain the content in perpetuity, they are responsible for finding someone else who can. While they may delegate some of their functions, such as ingestion or migration, to outside service providers, they are ultimately responsible to their clients for making sure these functions are carried out correctly and that the content is preserved for future access. This "integrated" responsibility for preserving and providing access to content may earn the trust of users, since the buck stops, as it were, with a definite party. Publishers, too, may be more willing to grant rights to a specific agent who takes responsibility for their work. Of course, the archive must live up to this responsibility, with its attendant certification and sustainability requirements. Penn, and most of the other Mellon-funded electronic journal archive planners, planned to build this sort of archive.

Distributed responsibility. The LOCKSS (Lots of Copies Keep Stuff Safe) project at Stanford suggests a more distributed form of responsibility, similar to the distributed responsibility for print journals discussed earlier. With LOCKSS, no one institution takes responsibility for a journal. Instead, many institutions maintain sites that cache copies of journal content published on Web sites. Content can be accessed from the cache if (and only if) it is no longer available on the original publisher's Web site. Cache contents are automatically checked against each other on a regular basis, and corrupted or lost copies of items replaced with other copies. The cache correction protocol is slow and relies on records of past holdings. These features make it highly unlikely, given enough participants, that a copy of an object will be completely lost, be corrupted, or become unavailable to any site that originally cached the content. Because the protocol also

checks what content a site has previously demonstrated that it owned, the protocol does not "leak" content to sites that were not authorized to see it.[13]

LOCKSS is designed to run on low-cost machines with as little maintenance as possible, making it attractive for libraries to set up LOCKSS caches. If enough libraries continue to cache the same content, that content (at least in its original format) can be preserved reliably even if no single institution takes final responsibility for it. The system can be queried to see how many sites are caching content. Institutions that see a need for greater reliability, can arrange to put additional copies on caches run at their own site or other sites.

In its original form, LOCKSS makes a number of simplifying assumptions about what is being cached, that may not be applicable to all journal archives. For instance, LOCKSS obtains publishers' content by automatically polling their Web sites for HTML and image files, but while those sites may include presentation forms of journals they typically do not include source forms. However, the basic LOCKSS model can be generalized to cover a broader range of journal archiving strategies, especially if caching sites can individually decide what to cache and how. Useful generalizations of the model include:

- **Introducing an explicit representation of trusted sources for journal content.** The basic LOCKSS model assumes the publisher's Web site is the trusted source for journal content. However, other trusted sources might be needed to provide content not available on the Web site such as source forms for journal content, metadata, and updates and migrated versions of journal content. Trusted sources could include publishers or well-known, responsible archives and service providers.

- **Caching Archival Information Packages (AIPs).** If AIPs and not just Web pages and images are cached, then a distributed archive can reliably cache both source and presentation forms of content, as well as appropriate metadata and migrated forms. These AIPs would need to come from a trusted source.

- **Identifiers for AIPs and other journal content**. A distributed archiving system requires a consistent way of identifying content so that different servers can check consistency between different copies of the same material. As originally designed, LOCKSS simply uses URLs as identifiers, but this only works for material pulled directly from the Web at static URLs. A distributed archiving system could use AIP identifiers to identify content, but would have to settle on a suitable global scheme for assigning such identifiers.[14] Other globally addressable entities, such as directories, will need globally unique identifiers as well.

- **Metadata.** A distributed archiving system should agree on a common core set of metadata to maintain for journal content, including descriptive and technical metadata, to support searching and browsing and maintenance of particular journal content.

- **Directory services.** Once the metadata above is in place it should be possible for archive users to be able to see what content is in the system, even if they cannot actually view all the content. They should also be able to search and browse it in customary ways, such as looking up an article by its citation, or browsing tables of contents for a particular journal volume. Directory information needs to be updateable as more journal content is added.

- **Versioning.** LOCKSS' cache consistency protocol works well for objects that do not change once they are imported into the system,[15] but a general-purpose distributed archiving system includes several types of changing objects. AIPs for particular journal articles may change as new manifestations of the articles are added (such as through migration) or if publishers issue a corrected version of an article after its original publication. Directories change as new articles, issues, and journals are added to the system. A versioning discipline would allow LOCKSS-like consistency checks for updates and new manifestations of these objects. A new version could be identified as an update of an existing version of an object in the system, and distributed archive sites could decide whether or not to accept the version. The sites could use acceptance criteria of their choosing, such as the trustworthiness of the source and the nature of the differences between the new version and the older version. Some types of updates may be expected to be monotonic: new data gets added, but old data does not get taken away. Sites could also decide whether or not to retain older versions.[16] Most legitimate updates to journals would not occur more frequently than issues of journals are published, a pace which is compatible with the deliberately slow coherency protocol that LOCKSS uses.

- **Controlled expansion of access.** LOCKSS' access control model is also monotonic: if a site has had a copy of some content in the past, it can be given a copy again, if its old copy has been lost or corrupted. Trusted sources could also grant new copies to sites that previously did not have copies of particular content, when this is permitted, without any further changes in the access control model. It would also be useful to be able to designate that certain content could be given to a wider set of sites. The simplest approach would be to include an "open access" flag that could be given to journal content or metadata that can be distributed freely. Trusted sources could set this flag, either on original distribution, or when previously arranged "triggers" applied that opened access to content.

In making these extensions, one must avoid introducing so much overhead into the system that the primary advantages of LOCKSS — reliable distributed archiving with minimal cost and maintenance — are lost. Each of the enhancements mentioned above should not be overly burdensome in itself, but designers of a distributed archiving system should make sure the enhancements are kept as simple as possible.

Service providers. Service providers perform specific functions on behalf of the archival system. In an "integrated responsibility" model, service providers are essentially subcontractors for the main archive. In the "distributed responsibility" model, service providers are trusted sources, as defined above. Service providers help scale up an

archival system by spreading out responsibility for archival tasks. They can be useful when an archival function requires specialized expertise or resources. For instance, if a company provides low-cost, highly replicated reliable data storage, an archive might outsource backups to that company rather than maintain backups itself. Or, if content is submitted or stored in a variety of protocols and formats, ingestion and migration of this content might be usefully parceled out to service providers, each specializing in a particular format or data provider, instead of trying to have one organization handle everything.

Service providers often do not need to make the same sort of long-term commitments as archives themselves do, unless their service includes long term information storage. However, they still need to be visible in the design and operations of an archiving system. Archive users may want to make sure they are certified for carrying out their service correctly. Publisher agreements may need to allow them to access copyrighted material, and publishers may want to be assured that they do not distribute or alter this material without authorization. The business model for the archive or archival system also needs to include some sort of compensation for the services of the providers.

Registries. Several organizations, including JSTOR, Highwire, PubMedCentral, and various publishers and national libraries, are currently archiving journal material. We expect that other archival initiatives will also arise as a result of the Mellon-funded planning process. As the population of archives and archived materials grows, it will be increasingly important for archive maintainers and users to be able to keep track of who is doing what and with what content. A registry of journal archiving activity would make this possible. It would allow users to find content they needed, and allow archive maintainers to find service providers, look up technical information on archiving formats and practices, and track redundant archiving of journal content.

Several architectures are possible for such a registry. A centralized knowledge base could be set up to accept information about archival activities. The Jointly Administered Knowledge Environment (Jake) is already set up to record journal information in this manner, though it does not currently record archiving information for journals.[17] Peer-to-peer systems can also act as decentralized "registries." The Typed Object Model, for instance, propagates information about data formats and their conversions between peer "type brokers."[18]

The Open Archives Initiative (OAI) suggests a particularly promising architecture for a registry.[19] Using OAI, repositories can expose metadata about their contents which can then be harvested incrementally by any interested service. The OAI protocol is lightweight and easy to implement and add on to existing systems. We have implemented it from scratch for selected digital collections here at Penn, using existing catalogs and databases combined with about 500 lines of OAI-specific Perl code. Journal archives could be designed to expose metadata about their contents, and their archiving strategies and policies. Then, one or more archival information sites could harvest information from appropriately certified archives, and allow archive users and maintainers to search and browse the aggregated metadata. Setting up such a system would require some

agreements on "core" metadata that the archives would export, but it would be in the interest of most archives to agree on such metadata. The aggregated information site (what would appear to most users as the master registry) would not need to be particularly expensive to maintain. It would be automatically updated from the certified archives, and the software to browse and search its records would only need to handle the common "core" metadata format.

A journal archive registry should include at least the following information:

- Information on journals being archived including
    - their names and identifiers
    - the archives or archival systems preserving them, the scope of the issues and articles preserved by these archives, and the formats being archived
    - who is authorized to access the content
- Information on the archives and service providers themselves, e.g.,
    - For "integrated responsibility" archives, information on their certification including how they are certified, when they were last certified, and relevant reports from the certifiers
    - For "distributed responsibility" archival systems, information on how to find out the archival status of journal content, including how to locate and check the reliability and redundancy of the storage of particular journal content

Archival rights and responsibilities

An electronic journal archive has responsibilities to scholars and their institutions and to publishers. The archive is responsible for ensuring scholars can access journal content for as long as it is of scholarly value in a form that allows it to be used effectively for research, teaching, and learning. The archive is responsible for respecting the copyrights of authors and publishers in its stewardship of electronic journal content. It is also responsible for helping maintain a healthy climate of scholarship and scholarly communication. It should neither hinder publishers or libraries with overly burdensome procedural and economic constraints, nor hinder scholars themselves with overly restrictive policies for access or deposit. Archives need to be granted sufficient rights to carry out these responsibilities.

In this section, we summarize the specific rights and responsibilities we believe archives need to have. The content of this section is based in part on our discussions with our publishing partners. Note, however, that these publishers have not made any binding agreement to these recommendations. Since we do not currently plan to archive their journals ourselves, binding agreements would ultimately need to be made with the archive or consortium that does archive the journals. We at Penn can, however, work with both the archives and publishers to negotiate an appropriate agreement.

Responsibilities for selection. The archive is responsible for identifying journals it wishes to archive and making these selections known to a publisher for the publisher's

approval. The selection does not have to be enumerated. For instance, archives that plan to include all electronic journals of particular publishers, as we originally planned, could agree that any newly added journal will be included in the selection unless either the publisher or the archive notifies the other that the new journal should be excluded. In such cases, the publisher should also inform the archive about newly added electronic journals or of journals that it no longer publishes. Of course, this notification, and other provisions of information given in this section, can be made simply by publishing the information in an agreed-upon location, such as a Web page or a mailing list.

The archive and publisher also need to agree on what content in the journals is being selected. Normally, the archive should collect at least all issues published from the time the agreement is made. Back issues, where feasible, are also desirable to include. As mentioned earlier, we expect that a scholarly journal archive should include at minimum all editorial content of the journal, including text and images, in some format. In our proposals to publishers, we specified that we would not guarantee preservation of "advertising matter, ephemeral and informal indexes (such as a page with links to current and past journal articles), search engines or other server programs, or auxiliary material not clearly part of the journal itself." Dale Flecker of Harvard has also recommended that information about the editing of the journal, such as the editorial board membership, the masthead, and author's submission guidelines, also be collected on a regular basis and preserved because of its potential interest to scholars analyzing the journal.

The archive and the publisher also have to agree on a set of formats that will be collected, and protocols for collecting content and metadata. Either the archive or the publisher should be able to request, with sufficient advance notice, that archiving for a journal be cut off: that is, no further issues will be deposited into the archive. The archive is responsible for notifying its clients, including appropriate registries, of its selections.

Responsibilities for ingestion. The publisher is responsible for providing content and metadata for the journal issues being archived. Content should include the journal content as provided to subscribers in presentation forms such as PDF and HTML. Even if an archive does not store presentation forms, as we recommend, it is helpful for an archive to be able to compare any source forms it receives against what the publisher actually supplied to subscribers in the electronic journal. This provision may simply consist of allowing the archive access to the online journal for testing.

When the archive is preserving source forms of content, the publisher is responsible for providing them and for ensuring they are correctly formatted; the publisher is also responsible for ensuring the source forms correspond appropriately to the published presentation forms. The publisher should likewise be responsible for providing metadata for the journal content in sufficient detail to allow standard scholarly citations to be built for all articles and other citable contributions. This includes title, authors, issue information, page numbers or other section delimiters where applicable, and other standard identifiers for the content such as ISSNs and DOIs. Abstracts and keywords would also be useful, when available.

The archive is responsible for collecting content and metadata in a timely fashion, checking it for consistency and proper formatting, and either ingesting it into its repository or informing the publisher of any errors or problems with the data. If the archive reports errors the publisher should remedy them, and the archive should then reingest content. Generally, the archive should try to ingest an issue while it is still the current issue. If the archive is harvesting data from the Web or another published source, it should be able to assume the data is in its final form unless the publisher has notified it otherwise. For example, if the publisher posts "draft" or pre-publication material which it then modifies before an issue's "official" publication, the publisher is responsible for telling the archive when and when not to harvest this information.

The archive is responsible for collecting appendices to articles such as datasets, audiovisual clips, program code, and other multimedia when this is within the archive's stated selection criteria. The publisher may need to assist the archive in collecting these appendices. If the appendices have unusual size or format, they might not be required to be migrated as is regular content. "Appendices" external to the journal itself — a Web site referred to by a URL in a journal article, for example — need not be collected.

If content is to be collected by harvesting from the publisher's site, the publisher is responsible for allowing access by the archive's harvesters, while the archive is responsible for ensuring the harvesters do not unduly load the archive's servers.

Rights and responsibilities for storage and maintenance. The archive is responsible for ensuring the long-term persistence and stability of the archived content. It is also responsible for making sure the content does not become unusable due to technological obsolescence. Therefore, the publisher should give the archive the right to store copies of the journal content and metadata, make and store backup copies, and create derivative works based on the original data for the purpose of maintaining their suitability for research and scholarship. Such derivative works could include migrations to new formats, indexing for searching and browsing, and transformations required for emulation. Ideally, the right to create derivative works would also allow enhancements to the content, at least when the enhancements are for services that researchers come to expect in online journals. Such enhancements, for instance, could include converting static images to wavelet-based forms that allow panning and zooming, or inserting hyperlinks to make it possible for readers to go to referenced articles.

Users expect that reliable, persistent electronic archives will not lose content. Therefore, the rights above and other intellectual property rights given to the archive should be irrevocable by the publisher, provided that the archive fulfills its responsibilities. Publishers can retain copyright; they just need to assign appropriate rights for the archive to do its job.

Rights and responsibilities for access and distribution. The archive is responsible to its clients for providing authorized users with its content and metadata, and it is responsible to publishers for providing content only to authorized users. Content provided to authorized users should include

- Copies of the journal content as originally published (byte-for-byte copy, in original formats) for journals archived in presentation form.

- Images of the journal pages in formats commonly readable at the time of access for journals archived in presentation form. This may require migration if common data formats change.

- The text of the journal content in formats commonly readable at the time of access. This may also require migration.

- Where feasible and cost-effective, other journal content, such as data sets and audiovisual materials, in formats commonly readable at the time of access.

Metadata should include at least citation information for the journals and their articles. The archive should also provide reasonable facilities for locating an article or journal with a known citation.

The difficult issue for negotiations in this area concerns who should be "authorized users" for content and metadata and under what circumstances. Indeed, this was a sticky point not fully resolved in our own publisher negotiations.

It seems relatively uncontroversial to allow general access at least to citation-level descriptive metadata. This allows users to know what is in an archive. Richer descriptions, such as abstracts, may be more controversial, but if they encourage a user to seek out the full content and become more interested in the journal as a whole, providing such detailed metadata can benefit both archive and publisher.

Likewise, allowing access to content by the publisher and by the archive maintainers seems uncontroversial. Widening access further, though, proved more problematic. Even providing access to subscribers to the journal for content that was also available on the publisher's Web site raised concerns that the archive would compete with the publisher's own offerings. The concerns here seemed to be not just possible loss of marketing opportunities on the publisher's own Web site, but also possible degradation of image. The look, the services, and the authority that the publisher took pains to establish on their Web site could be lost on the archive site, thus weakening the "brand" or reputation of the publisher. This seemed to be less of a concern if the archive just offered material that was no longer available on the publisher's own Web site. Concerns about image may also be alleviated if appropriate links (and distinctions) are made between the archive site and the publisher's site.

Even more controversial is allowing access to archived journal content to those who are not subscribers to the journal. JSTOR and Highwire, two established projects that archive journal content, allow such access after a certain time has passed from publication. The point at which access is opened is referred to as the "moving wall." There are several advantages to this sort of access:

- It satisfies, at least in part, the desire of many scholars that scholarly information should be generally accessible to all for little or no cost, whenever feasible.

- It provides an electronic counterpart to the long-standing custom of researchers obtaining older journal articles by interlibrary loan if their institution did not subscribe to the journal or retain their volumes.

- It avoids the overhead of the archive having to keep track of exactly what users should be allowed to access each item in the archive.

- It allows a wide audience of readers and third-party automated programs to examine the archived content and verify that it is being accurately preserved or report problems if it is not, thus enhancing trust in the archive. This may be especially important when material needs to be migrated to a new format, a process that may risk losing information or usability in unexpected ways.

- It may make it easier for mirror sites and other service providers to access the content.

- It provides an ongoing service of document delivery to the scholarly community, instead of just being an unseen "insurance" policy as an inaccessible or "dark" archive would be. This strengthens support for the archive in that community.

These are all benefits for the archive and its users, but they are not direct benefits to publishers. Of greatest concern was the possibility that libraries would cancel subscriptions to their journals or not sign up for them in the first place if they could have access to the archive without a journal subscription. Martin Richardson of Oxford University Press, for instance, reported that institutional subscriptions to eleven journals made available via Highwire had declined by three percent per year in the three years since they had been placed online, even though they had been increasing prior to that time. (The moving wall for free access to these journals ranged from twelve to twenty-four months after publication.) Faced with such declines in subscriptions, publishers might have to raise prices for the subscribers that are left, or cut the budgets to produce the journals. As an alternative to general free access Richardson recommended free access to readers in developing countries and low article pricing for infrequent users of journals.[20]

Richardson's report does not prove that opening access caused the drop in subscriptions or that a more distant moving wall, such as the three to five years that is common for JSTOR journals, would not solve the problem. Both Oxford and Cambridge have been willing to experiment with allowing some of their journal content to be accessible to nonsubscribers through JSTOR and Highwire as well as PubMedCentral. We recommend that electronic journal archives conduct further experiments to determine the relationship between events that would "trigger" nonsubscriber access rights versus paid subscriptions, since the potential benefits of general access to the archive and its users are substantial. Some may argue that copyright law already specifies a moving wall when works enter the public domain and are available to anyone. However, this moving wall

has been extended repeatedly in the United States and is now up to ninety-five years after publication for older publications and works for hire, and to seventy years after the death of the last surviving author for other works. The distance and uncertainty of copyright expiration in these circumstances makes an archive that requires public domain verification before providing access to electronic journal content equivalent to a "dark" archive for most practical purposes.

Other "trigger" events besides a fixed moving wall are also worth considering. For example, opening up access when migration is required would ease user concerns that migration might not be successful. It might also benefit publishers if they found migration too expensive or troublesome to carry out themselves. It is also reasonable for archives to be able to provide general access to content that is no longer being offered online by the publisher or the publisher's agents.

The archiving agreement should not abridge the traditional rights recognized by copyright law. These rights include fair use, first-sale rights, and unrestricted use of public domain material. All of these rights play important roles in scholarship, so archives whose purpose is to support scholarship should avoid restricting them. Some publishers digitizing older material may want to make an exception to this rule for new digitizations of public domain content from early in a journal's run. An archive should weigh this proposed exception carefully. In some cases, a short-term embargo on unrestricted use of this material may be necessary if the only likely digitizer of the content is worried about the investment going to waste, but there is little justification for extending this embargo beyond the same moving wall term (with the term in this case starting at the time of digitization) that is given to copyrighted content. On the whole, the copyrighted content of most electronic journal archives will usually be much more valuable than recently-digitized public domain content. Users are also much more likely to trust and support a comprehensive and authorized archive of a journal's run than an unauthorized provider of part of the journal's run.

Rights and responsibilities for certification and evaluation. The archive is responsible for specifying a process for certification of its content and procedures, and for reporting the results of this certification to its clients, including its publisher partners. The RLG-OCLC Trusted Repositories paper notes two basic approaches to certification: an approach based on auditing and an approach based on standards and usage. For a "dark" archive, or an archival system that uses software that is not available for public inspection, certification will generally need to involve auditing, to satisfy the concerns of constituents who cannot examine the content or the software themselves. In a more open system, such as LOCKSS, each site can examine its own contents and the software it is using. In such cases, separate auditing is less important though it may be desirable for the designers of the system to have the code reviewed and certified by an outside party.

Publishers may want to know how the content they publish is being used in an archive as well as the cost of the archiving processes. It is reasonable for an archive to share aggregated usage information with publishers as long as the form of the information sharing protects the privacy of individual readers. If publishers are subsidizing the

archiving they also have a right to be informed of the costs involved. Archives and publishers should agree in advance on the data an archive will collect, what an archive can be expected to provide, and for how long an archive should retain usage data.

Responsibilities for sustainability. The archive is responsible for ensuring it has appropriate technology, procedures, and funding for it to continue archiving activities for as long as needed. We will discuss technology and procedures in the next section. Responsibility for funding, however, is related to the other rights and responsibilities of an archive.

Funding for an archive can come from various sources including:

- The organization running the archive (self-funding)
- External sponsors (e.g., private foundations, government grants, marketers)
- The publishers of the journals in the archive
- Users of the archive materials (e.g., individual researchers and libraries)

Self-funding is viable for most academic libraries if the archival system is sufficiently lightweight and inexpensive to run and provides commensurate benefits. We can run a LOCKSS server on a commodity PC, for example, and in its current form it requires very little maintenance. Even a more ambitious system, such as the distributed archiving system described earlier, could easily pay for itself if its cost were similar to the original LOCKSS system and it made it easier for us to move older journals to inexpensive remote storage.

Full-service "integrated responsibility" archives are much more heavyweight. According to discussions with its maintainers, PubMedCentral, run by the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH), requires seven full-time employees to run on an ongoing basis. The project archives fewer than thirty externally published journals, plus a few dozen very low volume journals published by NCBI, and makes extensive use of automation. PubMedCentral staff represent only a small fraction of the staff of the NCBI, itself a small division of NIH. Organizations like this, which have national mandates as collectors and providers of research information, should find it a relatively small stretch to maintain an archive of online scholarly journals in-house and to justify it in their budgets.

On the other hand, most university libraries, including Penn's, would consume significant portions of their current budgets and staff to run an electronic journal archive. Penn currently has fewer than seven FTE equivalents designing and maintaining its digital library systems. If Penn were to archive all of the hundreds of journals published by Oxford and Cambridge its costs are likely to be considerably higher than for PubMedCentral's archiving of a few dozen journals. The Penn Library's mandate is not primarily archiving for an external community. Hence, internal funding for an archive that mostly deals with external publishers and users and that has significant costs without

clear bounds is not a viable long-term funding model. External grants, while they can be tremendously helpful for research and development and for startup costs, are much harder to obtain for ongoing operations which will continue to be expensive. Advertising support, as seen in the recent dot-com boom-and-bust cycle, is also highly unstable, and in a university environment can create problematic entanglements with commercialization. Neither external grants nor advertising or marketing support, then, will provide the reliable ongoing funding that an archive needs.

Many publishers can fund an archive if the archive provides them sufficient benefit. The costs of funding will ultimately be passed on to someone else. For subscription-based journals, the price of subscriptions may account for the cost of funding the archive. Journals that charge authors for publication might build charges for preservation of the article into the author's fee. Journals published by scholarly societies may have those societies fund their preservation.

Costs for preserving journal material are ongoing but publishers may be unwilling or unable to pay for archiving the material indefinitely. One solution to this problem is to structure publisher funding as an endowment. When a publisher deposits a journal volume or issue in an archive, it would also include a one-time payment to endow the future archive needs of the deposited content. Income from this endowment would fund ongoing archiving, backup, mirroring, maintenance, and migration of that journal content in perpetuity. If the archive decided to transfer responsibility for the content to another institution or back to the publisher, the content's endowment would be transferred as well. If a publisher stopped supporting an archive, the endowments associated with previously deposited content could still fund the preservation of that content.

For the publisher and journal subscribers, the endowment would provide a guaranteed source of funds to ensure the ongoing availability of journal content. Funding would not be dependent on an archive's ongoing revenue stream. By guaranteeing long-term availability of journal content, an endowment-funded archive could ultimately save money both for the publisher and subscribers to the journal. Subscribers who can rely on preserved electronic copies may be willing to forgo print versions or keep them in less expensive storage, thereby lowering their costs. Hence, they may be willing to contribute to an endowment fund via their subscription fees. If the archive is sufficiently trustworthy and accessible, it may even be possible to discontinue print production which would lower costs and subscription fees for the publisher.

For the endowment model to work, however, costs have to be well-understood and kept under control. If the costs can be kept down to a small percentage of subscription revenue, the endowment model may be financially viable. For instance, a journal that costs $350 per volume, with 300 annual subscribers, brings in about $100,000 of subscription revenue. If endowment income were five percent per year, then each one percent surcharge for archiving would pay $50 per year of archiving costs for that volume. So if the true costs to the archive were $150 per year per volume, three percent of subscription revenues would need to go towards archiving.

Unfortunately, this may be an optimistic estimate of archiving costs, which are still uncertain and likely to change as archives develop and as technology changes. Long-term costs may be difficult to measure in the early stages of an archiving project, especially since it may be unclear what costs are startup costs and what costs are ongoing or recurrent, given the fluid nature of technology and archiving standards. Automation of as many tasks as possible may help, as long as the cost of computers continues to decrease while the cost of labor increases. The endowment model also works best when an archive's fixed costs of operation are low compared to costs that vary with the size of the archive, favoring larger archives. For similar reasons, high endowment requirements may also favor larger publishers which might be an undesirable side-effect for libraries already worried about over-consolidation of scholarly journal publishing in for-profit conglomerates. If scholarly archives are ultimately designed to meet the needs of scholars and scholarly institutions, a funding model that relies on publishers may produce a less useful archive than one that is funded more directly by scholars and their libraries.

Funding by users appears to be a viable alternative to endowments by publishers. JSTOR has found success through funding from libraries and other journal users. Libraries pay a one-time capital development fee and an annual subscription fee to access JSTOR content. Publishers are not charged for deposits into the archive. JSTOR has proven highly popular with libraries and their patrons. At Penn, where it is one of 1,300 subscription sites, JSTOR's usage is in the top ten percent of the 300 databases we provide access to, with up to 1,000 logins per month during peak times of the academic year. Our scholars value JSTOR not just because it preserves journals from possible loss but because it provides them for everyday access, including access to materials not available in print form at our libraries since JSTOR subscribers can retrieve articles past the moving wall for any journal in the archive, regardless of whether they originally subscribed to the journal or not. In effect, JSTOR's convenient access is the carrot that encourages libraries to pay for subscriptions that cover both access services and archiving. JSTOR's success with libraries, scholars, and publishers makes a powerful case for this type of funding model.

Of course, it is possible for journal archives to have multiple funding sources, but complicated funding models may produce more problems than they solve. In an early proposal to our publishing partners, we proposed that publishers would pay for the core archival functions via an endowment and users would pay for user-friendly access services via a subscription. However, this model, even though it theoretically divided responsibilities between two types of functions, proved too difficult for people to keep straight, even within our own library, without repeated reiteration. Even when explained, its two-stream funding model made it too easy to argue about which stream should be responsible for which costs. We were also warned by the developers of some content sites that "enhanced" access services could have unexpectedly large costs to develop and maintain.

It seems more promising to have one main stream of ongoing funding for an archive, perhaps with additional assessments of funds for particular, well-defined functions. Such assessments may actually improve the operation of the archive. For example, ingestion is

likely to be expensive for many electronic journals, given the diversity of their delivery formats and the need to verify submissions before archiving them. An archive faced with high ingestion costs could depend on users to fund its general operations and development, as JSTOR does, but charge publishers for ingesting their content unless their content came in a prepackaged, easily-validated form prescribed by the archive. In effect, the archive would offer to sell publishers a packaging service for their content. The publisher could pay for the archive to package their content, or pay some other service provider to do it, or do it on its own and avoid having to pay. In any case, the archive's own costs for ingestion would be reduced and publishers would have an incentive to provide their content in accordance with archival standards. Both of these developments would make the archive more reliable and sustainable.

Funding and control of access rights represent a tradeoff. Publishers can provide funds or they can provide access rights to users after a suitable delay. They are reluctant to provide both. Libraries and scholars, on the other hand, may be disinclined to fund an archive that does not give them access except in some far-distant or unforeseeable circumstances. Thus, archives funded primarily by publishers may tend to be "dark," whereas archives funded primarily by libraries and archives should be more open to access by their funders. We at the Penn library prefer the latter for an archive that we run or help support.

Responsibilities and rights for transfer and delegation. Archives must be ready for unforeseen changes in the scholarly or online community that may render the archive unsustainable or make it far preferable to archive content elsewhere. Therefore, agreements with the publisher should allow the transfer, if necessary, of the archival material and its associated rights to another party that agrees to assume the archive's responsibilities. The publisher, if it is still in existence, may want to stipulate some additional qualities of the new party, for instance, that it not be a commercial competitor.

Similarly, the agreement should allow third parties to act as service providers for delegated archival functions as long as the third parties do not distribute archival material or use it for any purpose other than to carry out legitimate archival functions on the archive's behalf. For example, integrated archives may need to have a third-party off-site mirror of the archive's contents in case of a disaster at the main archival site.

The archival life cycle

In this section we discuss our experiences and findings in respect to the archival process. The OAIS model, as mentioned earlier, breaks down the archiving process into distinct functional modules which manage archival information packages throughout their life cycles. The key modules defined by OAIS and some of their functions are:

- **Ingest.** This includes the archive's acquisition of journal content and metadata, validation of the content, and packaging for archival storage as Archive Information Packages (AIPs).

- **Archival storage.** This includes maintaining AIP data and ensuring the continued survival and integrity of binary representations of archival data.

- **Data management.** This includes, among other things, maintaining metadata on archived content, managing storage, and handling queries on stored data.

- **Administration.** Among other things, this includes migration.

- **Preservation planning.** This includes planning and implementing further development of the archival system, including migration planning.

- **Access.** This includes delivery of content both to end users and to other archives under appropriate access controls.

While we had originally hoped to construct a full prototype archive for testing, this proved to be impractical for us, particularly if it had implemented the entire OAIS framework. Instead, our implementation planning and coding concentrated on a few of the archival functions we saw as needing special attention.

Ingestion. One of our focuses in planning was ingestion of presentation content and metadata from Web sites. We wrote routines to harvest information from publisher Web sites. Two programmers coded the implementation and an Oracle database server was used for data collection. The purpose was to construct a modular prototype harvester that could collect metadata and links to content from different publisher Web sites, then deposit the metadata and links into a common data structure that could be browsed and accessed uniformly.

Web sites for most Oxford and Cambridge electronic journals followed one of a few particular templates, so a few harvesting modules and templates could retrieve metadata for a large number of titles.[21] The modules successfully retrieved citation-level metadata as well as links to the article content, which in the journals we looked at tended to be PDF, HTML, or both. Article and issue metadata, including pointers to content, was then entered into a relational database using Structured Query Language (SQL). Another module repackaged metadata for particular journals in XML forms and assigned persistent identifiers to the content. A Web interface was built that used the XML to allow hierarchical browsing of journal content online.

Size of the content varied greatly among journals. For example, PDFs for *American Law and Economics Review*, a semi-annual journal, took up little more than 2 MB of storage space per year, whereas PDFs for *Brain*, a monthly medical journal, took up closer to 100 MB of storage space per year.

Articles in PDF form were generally easy to harvest, usually as one PDF file per article, although in some instances the PDFs were page-oriented instead and included content from multiple or adjacent items in the same file.

Correctly harvesting HTML presentation forms was considerably more problematic. Many of the HTML-formatted articles were clearly generated from source files and included numerous automatically generated links, some of which led to content that was important to the article, such as high-resolution versions of illustrations, while others led to related material that was not part of the article itself. Many of these HTML articles were produced by the same organization, Highwire Press. We did not attempt to harvest the HTML articles in this case although careful coordination with the online publisher probably could have facilitated reliable extraction of the actual article content. Since Highwire generates these HTML displays from archived source files, it would probably have been easier to archive the source files. It might, however, still be useful to harvest the top-level HTML page of an article and temporarily store it for error checking.

Ingesting material into an archive, whether manually or automatically, is far from foolproof. When publishers manually send a journal issue to an archive they may accidentally leave out articles, send misformatted files, or even send the wrong files altogether. Similarly, our harvesters were vulnerable to small changes in the HTML generated by the Web sites and could misparse or miss metadata when changes were encountered.

Occasionally, articles published on Web sites included mistakes. An early spot-check of one mathematics journal article revealed a PDF file in which the last pages could not be read as the file was initially posted. Human checks may turn up some of these errors but can be expensive to perform.

We found that the harvesting process afforded several opportunities to collect "redundant" data which could help flag errors in ingestion. In many cases, "table of contents" information for journals is sent out to interested subscribers by email. We are already collecting such email notifications and hope to eventually parse their contents for a "virtual subscription" service we plan to provide to e-journal readers at Penn. The parsed versions of these email messages can be compared against the metadata harvested from the journal Web pages and any significant differences flagged for human investigation. Similarly, our harvesting modules can be used to selectively harvest back-issue content as well as newly published content. At some specified time after the PDF for an article was first harvested the system could attempt to harvest the same article again. Differences could indicate that the content has changed or that one of the harvest attempts did not complete successfully. Either way, the anomaly could be flagged for an archive maintainer to check. We have not yet determined how useful these redundant checks would prove in practice, but suspect that given reasonably consistent publishing formats and well-tested software they could help reduce error rates substantially for low costs relative to volume.

Migration. Left alone, archived content and metadata are not likely to stay usable forever. Preserving the integrity of the original digital forms of archived electronic journals is not difficult with proper planning and the right to make copies as needed for archiving. The strategies of generating checksums, making backups and offsite mirrors, regularly refreshing media, and conducting regularly scheduled consistency checks and

disaster drills are well-known in digital preservation and data security circles. However, more uncertainty exists about whether the digital forms can continue to be understood and used effectively as technologies and expectations change.

Data structures and formats based on simple, openly-specified, and widely-used standards are the easiest to preserve, and journal archives should encourage their use. The definition of XML, for instance, is straightforward and widely publicized. Even if XML is eventually replaced by some other standard for structured data, it would not be difficult to migrate XML files to the new standard or to maintain programs that continue to parse XML in its original form.

To understand and use archival files formatted in XML or SGML, though, it is not sufficient to be able to parse the markup; it is also necessary to know what the markup represents. In addition, we need to preserve the DTDs or schemas used, as well as the documentation for defining DTDs, schemas, or whatever syntactic convention next comes along to describe markup formats. And, above the level of DTDs or schemas, we need to document and preserve the semantics of the schemas: the meaning of the markup elements and how their composition turns a set of angle-bracketed words and sentences into a journal that speaks from mind to mind. The semantics may be expressed in English-language descriptions, stylesheets, a program that renders markup as a visual display like that of a Web browser, or all of these. Particularly since some of the syntax and semantics may be specific to the world of electronic journal publishing and archiving, the archival system will need to include a preservation mechanism for both the syntax and semantics of its data.

Preserving format information is not a new idea. The Multipurpose Internet Mail Extensions (MIME) registry maintained by the Internet Assigned Numbers Authority (IANA) preserves descriptions of data formats commonly used on the Internet, as well as standard methods for encoding, packaging, and transmitting them. The Typed Object Model (TOM), now maintained at Penn, adds semantic capabilities to a distributed format registry mechanism, supporting automated services like migration (with configurable degrees of fidelity),[22] format identification, and interpretation of data formats by remote servers. An ideal registry for journal archiving data structures would allow the combination of the prose descriptions and canonical authority of IANA's MIME registry, the computational power of TOM, and the expertise of digital librarians, archivists, and publishers. In conversations with other digital library projects, we have found that such a registry would be useful to several types of digital library applications. We are now investigating the design and support of such a registry which could be the basis for a data format management service provider in an electronic journal archiving architecture.

Preserving presentation forms presents some special problems: they are often complicated, defined by vendors (in practice, if not in theory), and dependent on external information in unexpected ways. Fortunately, the presentation forms of most electronic journals are designed to be viewable by ordinary Web browsers without esoteric plugins. The capabilities of Web browsers are well-documented, for the most part, so it should be possible to deal with the formats they handle.

Articles in the journals we examined were presented in HTML or PDF. HTML is a well-known standard, and variations from the standard in practice are effectively defined by the behavior of the two major browsers, Netscape and Internet Explorer (the latter having a near-overwhelming share of browsers at this writing). These variations are also well-documented. Images inlined in HTML articles represented image formats — such as GIF, JPEG, PNG, and TIFF — that are also well-documented and that are each supported by a variety of tools. The most difficult challenges in preserving HTML involve embedded scripting, which has been handled noticeably differently in different browsers and browser versions, and hyperlinks to other content. The data at the end of these links may or may not be journal content that needs to be preserved, and in some instances may be dynamically generated by a server that is not likely to persist for nearly as long as the electronic journal itself.

Adobe's PDF, though defined by a particular proprietor and more complicated than HTML, is also practical to migrate in many cases, as long as Adobe continues to publish its specification and provide tools for viewing and manipulating the format. We have published an *RLG DigiNews* article giving recommendations for preserving PDF, including migrations that preserve the image and the main text of PDF files.[23] Again, scripting and external links pose potential problems. An important external linking problem in PDF concerns fonts, which may not be embedded in a PDF document but only referenced. As long as the fonts are for standard ASCII, ISO, or Unicode character encodings, external fonts pose little problem; programs that render or migrate PDF files can substitute standard fonts if necessary, and the PDF will still have sufficient information for correctly sizing and spacing the text. Occasionally, though, special fonts may be used to encode nonstandard characters, such as one might find in complex mathematical equations or chemical formulas. These may be unintelligible unless the specific font is preserved. The variety of characters available in Unicode should make the use of idiosyncratic fonts unnecessary in most cases but not all publishers have fully transitioned to Unicode. If nonstandard character sets prove to create significant problems, it may be desirable to also preserve a registry of unusual fonts or to embed the fonts into the PDF file at ingestion time. Both strategies have the potential for conflicts with copyright law, although the law allows some of these conflicts to be avoided as long as the fonts are only copied or embedded for nonprofit archiving.

We recommend attempting to migrate presentation files to other presentation formats, such as page images and plain text, as soon as possible, even if the migrations are not immediately needed. Archives will then have the tools and procedures they need for migration ahead of time. Testing of these procedures, even if the results are discarded after testing, should help the archive control its costs and improve its reliability.

The static appearance of presentation files can also be preserved through printouts, as other digital preservation studies have observed.[24] However, printout-based preservation can get expensive, particularly if the originals are in full color, and it loses the extra information in the electronic version, including a machine-readable and machine-searchable form of the text. Particularly prudent and well-funded archives may want to consider printout backups, but as long as electronic journals are published in

conservatively-structured files with standard formats, full printed backups are probably not necessary. There may be a place, however, for selective printouts, triggered when an analyzer detects that a file has unusual characteristics that might make it hard to use or migrate optimally.

As with ingestion, preservation and migration of journal content can be simplified if publishers settle on standard forms to use in their source and presentation files. Archives and their users should encourage publishers to use standard forms when possible and help define such forms where necessary, such as the standard XML form for e-journal source files recommended by the Inera study. They should also make sure that "standard" formats and delivery mechanisms continue to be open and supportive of archiving functions. Archives should be much more wary of proprietary formats without openly published definitions (Microsoft Word, for example). If it is necessary to ingest such formats, archives should try to migrate the files to formats more suitable to archiving as soon as possible since the only tools available for this migration might be proprietary programs with a limited useful lifespan. The introduction of "Digital Rights Management" should also be critically scrutinized. Such capability, designed to protect copyrights by restricting copying and use, may severely impede many archival functions and may be very difficult, both technically and legally, to work around.

Finally, even when journals have been fully and successfully migrated to new forms, we recommend also keeping the forms originally ingested by the archiving system if it is not cost-prohibitive to do so. Scholars have long found value in going to the original form of a document, or as close to the original as possible, often for reasons the original preservers did not anticipate. For example, the location of line breaks may seem like trivial information, but it may be of value to scholars, and since most migration of nontrivial data formats typically lose some information, the location of line breaks may not survive a migration. Even if no one cares to directly study the original binary representations of an electronic journal, archivists may in the future find new ways to migrate or display content that are more reliable and preserve more information which they can then apply if the original forms have been preserved.

Summary and conclusions

By the time we reached the end of the planning project at Penn, we found that many of our initial assumptions had changed. When we embarked on the planning project, we hoped to act as a primary archive for most or all of the journals of two university-affiliated publishers. We initially focused on preserving presentation forms of journals, automation of as many functions as possible, pre-planned migration strategies, relatively open access to content after a few years, and verification of the content by its users, as well as by automated checks. Funding would come from publisher endowments and from users of advanced access services. By the end of the project, we hoped to have both a full working prototype archive and agreements with publishers for permanent archiving of the journals. We found that our reach exceeded our grasp but that the planning process and the problems we uncovered could help guide formation of viable archival communities.

In our planning year, we negotiated with Oxford and Cambridge for the archiving of their electronic journals. We also developed prototype code to harvest journal metadata and content from their Web sites to facilitate low-cost, automated archiving of presentation forms for the journals. We have not come to a final agreement with the publishers, in part because we do not now plan to be the primary archive for their journals as we had originally intended, and because the consortium-based archiving solution we hope will archive their content has not yet been established. However, both publishers are interested in having third parties archive their material under appropriate conditions and safeguards and have made agreements for third parties, such as Highwire and PubMedCentral, to archive and share at least some of their journals. If a viable consortial archiving arrangement does emerge, we would be happy to work with Oxford and Cambridge in negotiating their agreements with the consortium. We may also be able to assist in ingestion of their content by further developing and maintaining our code to harvest material from the Web.

Our discussions with publishers and peer libraries and archives were also illuminating. While we initially intended to focus on preserving presentation files, both publishers and other libraries made it clear that they found the extra structure and functionality of source files to be an important part of preserving the journal content. While archiving source files can be costly, particularly in the ingestion stage, they do appear to be desirable enough to archive along with presentation files. Standardization of source file format, such as recommended in the Harvard-commissioned Inera study, would help control these costs and allow more consistent levels of service in archive access. Source file standardization was also appealing to publishers.

Migration and other types of conversion will be necessary in the archive, and we continue to believe that pre-planning these migrations is important. Standardized XML formats for source files should be easy to migrate since XML is now ubiquitous. Migration of HTML files without extra scripting beyond standard stylesheet forms or server-side support should be accomplishable without difficulty as well. Migration of PDF, while more complex, also should remain feasible in all versions up to the current version (1.4), as long as the files are checked to ensure that fonts, scripting, and digital rights management

do not get in the way. We recommend retaining the original file format of archive submissions as well, if they are not unreasonably large, since we cannot always anticipate what information might be lost in a migration step that would be of interest to future scholars. Setting up distinct services to maintain information on archival data formats as well as conversions and other operations on those formats will help keep archive content usable as technology changes.

The experience of JSTOR shows that one organization can coordinate many of the core functions of an electronic journal archive. It also shows that given sufficient delay, a moving wall for opening access to older journal materials does not appear to jeopardize subscription revenues. Wide access to these materials both increases confidence in their reliability and provides a broad base of financial support to the archive from its subscribers. A critical mass of journals in the archive also helps gain subscriber support. An integrated journal archiving organization is more likely to be sustainable, at least in the short term, than a large number of smaller, independently run archival organizations, each creating redundant technology and systems and each having to compete for subscriber support.

In a JSTOR-like access model, funding by libraries and other users of an archive is more likely to be viable than heavy reliance on publisher funding. The endowment model we originally proposed for subsidizing archiving in perpetuity, even if feasible for some larger publishers, may be overly burdensome on small publishers and on new cooperative scholarly publishing ventures. Moreover, publishers that do fund all or most of the archiving are likely to want more restrictive access terms than many scholars and libraries want. Except for charges for specific services, such as initial publication or archival ingestion of journal content, it is best that integrated archives supported by the scholarly community also be funded by that community, that is, by the libraries and other users of the archive. That community should also have access to the contents of the archive after a suitable interval (in the neighborhood of five years) has passed from original publication.

A consortially-run archive is not likely to please everyone nor archive all of the journals that will be of interest to scholars. Therefore, development should also continue on lightweight, distributed archival technology like LOCKSS, which allows individual libraries and other journal subscribers to decide what they want to store, with increasing reliability of the content as more participants decide to store the content. When the technology is fully developed each library can retain the power to archive what it finds important, while being able to delegate archiving of established, widely-recognized journals to a central organization.

With several archival efforts by publishers, academic libraries, and national libraries already underway in various parts of the world, and with more informal distributed archival networks in development, the archival community will need some way of keeping track of what journals are being archived and by whom. It will also be useful to keep track of service providers for archival support functions including format and migration handling. A registry service could fill a needed role here. It could be set up in

conjunction with a consortial archive or it could augment existing or planned registry and directory services like Jake or TOM.

The Penn Library is ready to support a well-organized electronic scholarly journal archive that meets the needs of researchers here and at other universities. The success of the JSTOR project, to which Penn subscribes, shows the feasibility of providing substantial collections of journal runs to benefit scholars and the widespread support it can draw. We hope our experiences and recommendations can help make JSTOR-like archives of electronic content beneficial to scholars, cost-effective, and sustainable. We are also interested in participating in future tests of LOCKSS improvements, to advance the complementary model of distributed library-based archiving, and can dedicate some of the equipment acquired in our planning year towards this end. Finally, we are interested in the possibility of offering services to the archiving community such as ingestion of journal material and registry support for data format information and migration.

To summarize the next steps we recommend taking:

- We recommend following in the footsteps of JSTOR in setting up a library-supported organization for archiving electronic journals. We will encourage our publishing partners to submit electronic journals to a well-designed archive along these lines and can work with them to help set up appropriate agreements and ingestion procedures.

- We recommend further development of LOCKSS technology (with the enhancements recommended in this report) to support lightweight distributed library archiving as a complementary strategy to setting up an archival organization. We will set up a LOCKSS cache at this site so we can help test the technology and will suggest ways it can be optimized for reliable journal preservation.

- We recommend setting up service providers and registries to support archival functions and user needs. We will plan to provide such services ourselves. Initially, we plan to further develop TOM to better support data formats, migrations, and systems used for digital preservation, and to explore possible use of the technology as a basis for a community registry of data formats and services for preservation.

With several years' worth of electronically published scholarly journals now online and increasing dependence of scholars on the electronic medium, the time is riper than ever to move forward with electronic journal archives. We look forward to working with the community to help build and sustain them to ensure reliable, accessible, and robust scholarly communication.

Postscript: 2003

Since this report was written, the Penn Library has continued its involvement in technical initiatives to support long-term preservation of electronic journals and other digital information.

We are working with the Library of Congress-led National Digital Information Infrastructure and Preservation Program (NDIIPP) to help develop a distributed architecture for archiving digital information. The architecture is being designed to support a wide range of archival systems and strategies, and to allow institutions to collaborate in preserving, migrating, and exchanging digital information. The NDIIPP architecture group is designing tests of archival interoperation to be conducted next year, using an existing multimedia Web collection. More information on NDIIPP can be found at http://www.digitalpreservation.gov/.

We have also now joined the LOCKSS network, described in our report, and are helping integrate LOCKSS with proxy systems for everyday use of LOCKSS-cached content. The LOCKSS project is also working on archiving a wider variety of content, which we hope will test and improve its robustness as a general preservation system.

The need for long-term support of data formats has been a recurring theme in meetings of NDIIPP and other preservaion initiatives. Penn has recently received a grant from the Mellon Foundation to develop practical applications of TOM (which we described in the report) to support the handling of diverse data formats for digital preservation and online learning systems. We will be releasing TOM-based conversion services, format brokers, and program libraries this fall as open source software. With this software, publishers and archives can describe relevant data formats and provide identification, migration, and emulation services for them. We hope that this software will also support data format registries serving the digital preservation community. Updates and additional information on TOM and its services can be found at http://tom.library.upenn.edu/.

Our report found that documents in the PDF format were preservable if they conformed to certain constraints. Since then, international efforts have proposed a standard for restricted forms of PDF designed to facilitate preservation. We hope that this standard, known as PDF/A, will be widely adopted and aid in the preservation of PDF-based documents. More information on PDF/A can be found at http://www.aiim.org/pdf_a/.

Endnotes

[1] For more on the JSTOR model, see http://www.jstor.org/about/.

[2] For more on the LOCKSS model, see Stanford's report in this publication. See also http://www.lockss.org/.

[3] University of Pennsylvania Library, "2002 Quality/Impact Survey," Unpublished report (2002).

[4] University of Pennsylvania Library, "Usage statistics for July 2001-2002." The most popular e-journal, *Nature*, had 15,396 logins through the Penn Library in this timeframe.

[5] These migrations were infrequent, typically occurring only once in decades of use and bringing with them other opportunities to save costs: for instance, by freeing shelf space once taken up by full-sized print copies.

[6] From an analysis of http://www.jstor.org/about/issues/.

[7] 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. This document is also included elsewhere in this publication.

[8] RLG-OCLC, *Trusted Digital Repositories: Attributes and Responsibilities* (Mountain View, CA: Research Libraries Group, May 2002). Online at http://www.rlg.org/longterm/repositories.pdf.

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

[11] Budapest Open Access Initiative, "Statement of 14 February 2002." Online at http://www.soros.org/openaccess/read.shtml.

[12] *DSpace Federation*. Online at http://www.dspace.org/.

[13] Vicky Reich and David S. H. Rosenthal, "LOCKSS: A Permanent Web Publishing and Access System," *D-Lib Magazine* 7.6 (June 2001). Online at http://www.dlib.org/dlib/june01/reich/06reich.html.

[14] The OAIS model itself does not enforce unique identifiers beyond the scope of a particular archive, though federations can create globally unique identifiers.

[15] If a LOCKSS server changes the content of an object, but other servers do not, the changed copy is considered to be corrupt and is corrected to agree with the others.

[16] Note that sequences of monotonic updates require very little extra storage space for older versions, since only differences from the newer version need to be specified.

[17] Kimberly Parker, Cynthia Crooker, and Dan Chudnov, "Jake: Overview and Status Report," *Serials Review* 26.4 (2000): 12-17. Online at http://jake-db.org/docs/jake-overview-serrev.pdf.

[18] John Ockerbloom, "Mediating Among Diverse Data Formats," diss., Carnegie Mellon University. Technical Report CMU-CS-92-102 (1998). Available online in Postscript from http://tom.library.upenn.edu/pubs/index.html.

[19] The Open Archives Initiative (OAI) is a different initiative altogether from OAIS. See http://www.openarchives.org/.

[20] Martin Richardson, "Impacts of Free Access," Letter to *Nature* (5 April 2001). Online at http://www.nature.com/nature/debates/e-access/Articles/richardson.html.

[21] There were, however, some exceptions for older content and for journals with experimental features.

[22] Jeannette Wing and John Ockerbloom, "Respectful Type Converters," *IEEE Transations on Software Engineering* 26.7 (July 2000): 579-593.

[23] John Mark Ockerbloom, "Archiving and Preserving PDF Files," *RLG DigiNews* 5.1 (15 February 2001). Online at http://www.rlg.org/preserv/diginews/diginews5-1.html#feature2.

[24] For example, see Donald Waters and John Garrett, co-chairs, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Commission on Preservation and Access and the Research Libraries Group (1 May 1996): 27. Online at ftp://ftp.rlg.org/pub/archtf/final-report.pdf.