Project Harvest:
A Report of the Planning Grant For the Design of a Subject-Based Electronic Journal Repository

Presented to
The Andrew W. Mellon Foundation

Submitted by Sarah E. Thomas
Principal Investigator
Carl A. Kroch University Librarian
Cornell University

1 September 2002

Table of Contents

Introduction

In December 2000, in response to a call from the Mellon Foundation, the Cornell University Library received a grant to develop a plan for a repository of electronic journals in the field of agriculture. The Mellon Foundation recognized that solutions to the problem of preserving electronic journals can only be solved if done in cooperation with the publishers. From January 2001 through March 2002, the Cornell Mellon teamed worked together and with the project teams from the other Mellon planning grant recipients to deepen our knowledge of the digital archiving problems.

Project Harvest, as the project came to be known, built on Cornell's historic excellence in preservation in general and the preservation of agricultural literature in particular. During the course of the year we initiated a dialogue with a number of agriculture publishers with whom we have successfully cooperated on other projects. We sought to explore the conditions under which a publisher might be willing to participate in a subject-based repository. In addition, we surveyed specialists in the field of agricultural preservation in order to determine the requirements of librarians for digital archives. Finally, we spent much of the year exploring potential business models for a successful digital repository.

Cornell University Library's Interest in Digital Preservation

Cornell University Library has traditionally invested heavily in preservation of all kinds. The Preservation program is one of the best in the nation, with a staff of thirty involved in a range of functions from fine restoration (four professional conservators and ten conservation technicians) to digital preservation where a staff of five is devoted to research and applications. In addition, Cornell has a special mandate to preserve agricultural materials in relation with the National Agricultural Library and the United States Agricultural Information Network (USAIN). Cornell's interest in research resources covers a very broad spectrum. In addition, we are interested in serving both the immediate and long-term user, and have served as a *de facto* archive for content providers of all types.

In short, the Cornell University Library is deeply concerned with identifying and applying effective and efficient means for managing research resources, in avoiding redundancy/duplicative efforts, and in stabilizing materials to make them usable. This is easier when resources come on stable, eye-legible media such as animal skin, paper, palm leaf, even jade. It's more difficult when the medium contains the seeds of its own destruction, such as brittle paper, color transparencies, nitrate negatives. More modern media, such as videotape and sound recordings, have very short life expectancies. The problem is compounded when the media is dependent on a playback device which in turn may be subject to obsolescence. And, in the digital world, software dependency adds an additional layer of difficulty. The rate of obsolescence can be very fast — as short as a three-year window. Technological obsolescence is not the only problem in the digital world: more and more of the resources research libraries depend on are licensed, not

physically owned. A recent survey of Digital Library Federation (DLF) members indicated that 40 percent of their expense for building digital libraries goes to licenses.

Like other digital materials, e-journals are at risk from ongoing technical, organizational, and economic changes. For these digital assets to remain usable and valuable over time, there must an explicit, recognized commitment to maintaining the integrity of and ensuring the long-term preservation of e-journals. A digital archive has a key role to play in this digital life cycle by serving as a trusted third party for the preservation of digital materials; by establishing a secure repository that complies with accepted preservation policies, procedures and standards; by identifying or adapting improved and appropriate preservation practices; by supporting efficient, economical long-term access that balances the potential of developing technologies with available resources and required revenues, as appropriate; and by providing a reliable, monitored, maintainable infrastructure.

Preservation is also closely aligned with trust. The more control over a source document you have, the greater the ability to exert preservation measures. Research libraries have built in redundancies in their physical collections with a good portion of collection overlap. This would be difficult to duplicate in the digital realm because so much material is licensed and to replicate a digital archive at each site would be prohibitively expensive. Thus, the idea of trusted digital archives comes into play.

Skepticism remains strong among research libraries and their constituencies. Very few research libraries have withdrawn hardcopy versions of materials accessible in digital form. A recent survey by JSTOR of 4,220 faculty across the country revealed there is a growing dependency on electronic resources but continuing skepticism about their long-term viability. Nearly 78 percent of respondents indicated that hard copy versions should be retained even if an effective digital preservation strategy were in place, while 97 percent of respondents indicated it was important for libraries, publishes, and other partners to archive, catalog, and protect electronic journals.

Given this background, the challenge facing the Project Harvest team was to identify what was needed to foster digital preservation. Specifically, we sought to determine if the time was yet ripe for the following:

- Technical solutions that retained flexibility and some measure of reversibility.
- Cost-effective solutions based on sustainable business and organizational models.
- The establishment of third-party archives that would be trusted by publishers, users, and libraries. The goal would be for Cornell to archive agriculture e-journals in a way that would obviate the need for other research libraries to do so. Furthermore, scholars and others would trust the arrangement.
- The definition of an archiving solution that is verifiable and auditable.

The Nature of the Digital Archives

Levels of Access

Much of the first part of the year was devoted to an internal discussion of the nature of the digital archive that Cornell would be willing to maintain. Two potential models on a broad spectrum of possibilities were identified:

- "Dark" archive — A "dark" digital archive would be a closed repository that would strictly control individual and/or organizational access to the information stored under its control. Bits would be preserved in the event that the publisher no longer could provide access to the journal. The primary function of the archive would be as a fail-safe bit repository.

- "Light" archive — Conversely, a "light" archive would be a repository that would allow individual and/or organizational access to the information stored within. Access to the content of any of these options may be subject to access restrictions agreed upon by the publisher and the archive. Nevertheless, access under some circumstances would be presumed, and an access system would have to be maintained.

As we worked through business models in parallel with our discussion on the level of appropriate access, we came to see that the response to these issues would drive the design and organization of the entire repository. Our initial analysis, for example, suggested that a dark archive would be less expensive to build and maintain, but it also removed any potential short-term sources of funding. The contents of the archive would only become of value if the material were no longer available from the publisher. A light archive might be able to sustain itself as a secondary means of access to the content. In addition, regular access to the content in the archive would ensure the material was still usable. (A dark archive, conversely, would need elaborate systems to ensure bit integrity was maintained.) The light archive, however, would have much higher development and maintenance costs since in addition to storing and migrating data as with the dark archive, an access, retrieval, and authentication system would have to be maintained.

In the end, we concluded that Cornell should consider maintaining a dark archive when an appropriate business case can be made — namely, when someone is willing to subsidize the costs associated with maintaining a bit archive. The bulk of the Cornell University Library's efforts, however, should be devoted to developing and sustaining a light archive for public access with operational parameters to be set through discussion with its publishing partners. The nature and degree of access will have to be specified in agreements with our publisher partners and users.

Factors to Consider in the Development of an Accessible Archive

CUL believes the following parameters are significant:

- The user must have clearly defined use of the intellectual content of an electronic journal.
- Time constraints to access should be lifted with a "moving wall" similar to the JSTOR model.
- All information must be searchable by common metadata terms: author, title, publication, keyword, etc.
- Information retrieval should be defined at a common granular level. At the least, a user should be able to search and retrieve information at the article level. Also, the user must be able to browse different levels of aggregation including article, issue, volume, and title.
- Access must be assured following changes in publishing organization ownership.

The Subject-Based Digital Archives Approach

The course of this discussion led two team members to develop the distinctions further in a draft white paper on the Subject-Based Digital Archives (SBDA) Approach. A preliminary report on this analysis was presented at the Fall 2001 DLF Forum.

Most of the other Mellon project recipients did their planning around building archives for specific publishers. The Publisher-Based Digital Archives (PBDA) approach focuses on the distinctiveness of each publisher/journal. The SBDA approach, however, would stress the commonality of content across the full publishing spectrum (and beyond). Commonality supports and encourages access, increased access promotes buy-in from content controllers and creators, buy-in further increases access, and revenue from access can be fed back into preservation. One way it does this is by extending the "dark/light" dichotomy to metadata as well. It recognizes that one can have dark metadata with dark data; light metadata with light data; light metadata with dark data; and light metadata with no data. The actual scenario will depend upon the publisher. In certain cases, the ability to search the light metadata alone may generate enough revenue to support the maintenance of the archival system. The SBDA scenario may have important implications for preservation as well.

Developed as a straw man during the course of the project, the SBDA scenario held enough promise to warrant further work. The idea is being further developed in a report for the Council on Library Information Resources (CLIR).

Business Model Development

Ongoing funding for any digital archive must be predictable and also flexible enough to address future changes within the partnership. Early on in the project we recognized that CUL and its partners would have to take an active role in establishing alternative funding sources which might include access fees, grant funding, and endowment support.

As the project evolved, it became clearer to us that the development and maintenance of a digital repository that would meet the requirements of developing archival standards — including the OAIS Reference Model[1] and the RLG-OCLC report, *Trusted Digital Repositories*[2] — would be expensive. The design and preparation of the system would be one of many costs. As we learned from our partners working on Project Euclid, a math journal publishing project, the ingest of complex digital objects requires a degree of manual oversight and processing. In the absence of an acceptable technical model for an archive, however, it became impossible to accurately determine cost models.

The development of technical solutions for the archive was an essential prerequisite for our business planning. The technical model itself, however, needed to be shaped by business needs: the best technical model in the world would not be acceptable if there was not a business plan that could support it. This chicken-and-egg conundrum was one the project never successfully solved.

Metadata

Metadata is a broadly used term. "Descriptive" metadata records the content of a digital object, "structural" metadata records the structural information about a data object, and "administrative" metadata records the maintenance of the digital object. Users, archive managers, and archive auditors require metadata of all three kinds. CUL and the partners will need to share metadata for the archive via a common interpretation of an established standard. It became apparent to us during the course of the project that these metadata protocols will need to be implemented at Cornell in collaboration with other electronic journal projects such as Project Euclid. As necessary, the archive will modify metadata for local implementation, which may supersede proprietary metadata. In the long term, metadata costs for the archive must be minimized, and it would be expected that the publisher partners would efficiently accommodate metadata modifications adopted by the archival community. The project would adopt content creation policies to capture requisite metadata.

Preservation formats

It is possible the archive could store two formats: a publisher's proprietary format for typesetting or publication and an archival format with the emphasis on intellectual content. CUL, in consultation with its partners and other archives, came to believe ultimately that an acceptable archival format is highly desirable. A common format should encourage uniform archiving protocols and reduce the administrative overhead of the archive. Ultimately, it is expected the partners will submit files to the repository

already in a preservation format, reducing costs and errors associated with data conversion. The archive will provide a reasonable time for partners to achieve a coordination of formats with a goal of three years. One of the most successful parts of the planning year grant, therefore, was the dialogue that was started with Harvard and the National Library of Medicine on the design of an Archival Information Package (AIP).

Work with Publishers

With these general guidelines in mind, the project then turned to a group of publishers to determine their perspectives on third-party subject-based archives. A meeting with a group of publishers was held in Washington, D.C. in September 2001. Representatives from the American Dairy Science Association, Academic/Elsevier, the American Phytopathological Society, BioOne, CABI, NRC-Canada, Wiley, the National Agricultural Library, and USAIN met with members of the Project Harvest team to discuss the issues the team had investigated on its own.

At the meeting we identified a number of incentives that might encourage a publisher to arrange for the maintenance of its journals in a third-party repository. These included:

- Protection of assets, especially if the material has continuing value as it ages
- Low additional overhead for the publisher
- Customer satisfaction
- Potential advertisement for their materials

At the meeting we learned that all the publishers in attendance intend to establish their own archives. They saw themselves shifting from focusing on the currency of their content to developing databases of content of continuing value. Retrospective runs of journals which in the past they had been happy to leave in the hands of libraries become instead a potential source of new revenues. Much of the discussion centered on exactly what needed to be archived, and it became apparent that the publishers by and large were much less concerned about preserving the "artifactual" nature of the electronic document than about ensuring that meaningful content is carried forward.

It was clear during the course of the meeting that the publishers and the librarians in attendance had different perceptions concerning who should be responsible for digital preservation. Librarians, as the survey (Appendix E) revealed, want trusted third-party archiving. The publishers seemed unaware that some of their customers do not believe that the publishers alone safeguard materials.

Given their assumption that they would be archiving material in order to support their own revenue streams, publishers saw little need to pay to support a third-party archive. Likewise, given their interest in potential new revenue streams from retrospective

holdings, the publishers were not enthusiastic about "light" archives. A few would consider the possibility if revenue generated was returned to the publisher.

The good news was that on a technical level there appeared to be a real convergence in formats, with all of the publishers moving to an SGML-based publishing system. Many were unwilling to share the Document Type Definition (DTD) that they use — in some cases because of anti-trust concerns — but all seemed willing to consider developing as an output from their system an AIP- or SIP-formatted document[3] — assuming we can come to some sort of agreement about what each would contain.

An important part of all discussions of dark archives is consideration of what trigger events might move content from a dark archive into the open. The publishers were unable to come to any common agreement over what might constitute a trigger event. Some acknowledged that the passage of time might be one such trigger event, but they were thinking in terms of centuries, not the relatively short periods that are normally discussed.

Librarians' Perceptions

It became clear to the team early on in the project that if were to develop a repository that was to be trusted by other librarians and scholars, we would need to know more about what that community expected from such an archive. We therefore conducted a survey of preservation officers at USAIN and Land Grant institutions. The survey form is found in Appendix E.

The results of the survey were most revealing. Among the findings were:

- 45 percent of respondents indicated the need for both print and electronic copies of journals
- 55 percent of respondents indicated that e-journal already substitute for print
- 84 percent of respondents would cancel print if a trustworthy and reliable archive existed

When asked if they had detected a difference in content between print and electronic journals, 22 percent said they had noticed a difference, an equal percent said they had not noticed a difference, and 45 percent said they did not know. As for what a trusted repository should preserve, most of the respondents wanted the archive to maintain the "look and feel" of the journal as well as all the functionality that the publisher offered, while a smaller group would be happy with just maintaining the "look and feel." Most importantly, over 90 percent rejected any single archiving solution, preferring instead that multiple custodians or a third party do the work.

Conclusions

At the end of the planning year, Cornell University staff have a much clearer sense of our own expectations of what will be required in a digital electronic journal repository. The important work accomplished during this first year in translating the OAIS Reference Model, RLG-OCLC's *Trusted Digital Repositories: Attributes and Responsibilities*, and the various emerging preservation metadata standards into the Cornell environment continues in two important areas.

First, much of the Project Harvest work is being translated to Project Euclid (http://projecteuclid.org) and its newest iteration, the Electronic Mathematical Archiving Network Initiative (EMANI at http://www.emani.org/), an international collaboration for the preservation of the journal literature in mathematics. Several compelling arguments developed during the course of Project Harvest have led us to build the Euclid infrastructure. Though several options exist, we have decided that a subject-based archive can best be built around the *article* rather than the journal issue. Project Euclid is built around the journal article and therefore lends itself to this sort of approach.[4] Further, Euclid's modular component infrastructure as well as its support for OAI will make it possible for us to include in the system items other than journal articles, including gray literature, technical reports, and other items that would be appropriate for a subject-based archive. However, since Project Euclid was developed as a publishing system and not an archiving system, we will need to add to its infrastructure those elements that will allow the system to manage and maintain archival information packages as part of the system. We will therefore employ input from preservation policy staff and programmers trained during the course of Project Harvest to add the component parts to the existing system to make Project Euclid an archival (as opposed to publishing) system compliant with OAIS.

While we are excited about the development of the EMANI project, the Project Harvest planning process also raised real issues in our minds about the viability of managing national, and even international, electronic journal repositories in individual institutions. We were fairly certain by the end of the project we could develop a viable technical infrastructure for the repository. It was far from clear, however, that we could develop a funding model that would sustain that repository. Publishing partners were reluctant to either fund directly or indirectly (e.g., through higher subscription costs) the maintenance of such an archive; early investigations of a subscription model among potential archive clients, while promising, still faced the challenge of "free riders;" and the responsibility for maintaining a repository for a discipline is something that no institution should have to take on alone. Further work on the SBDA model may lead to the conclusion that it could become a reliable source of revenue for the archive. At the last meeting of the Mellon participants, however, our attention shifted to the planning process for the development of a central archiving service. The recognition among the Mellon e-journal archive planning participants that the function is best performed centrally may be the most important conclusion of all.

Endnotes

[1] 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.

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

[3] The OAIS reference model describes the organization of digital content into "information packages," namely, Submission Information Packages (SIPs), Archive Information Packages (AIPs), and Dissemination Information Packages (DIPs).

[4] We will leave it to publisher-based archives to concentrate on the publisher's output as defined by the publisher, and to address issues related to the capture of the "look and feel" of the journal issue.

Appendix: Project Harvest Team Members

Project Harvest Team

Sarah Thomas
University Librarian
Principal Investigator

Peter B. Hirtle
Director, Cornell Institute for Digital Collections
Project Coordinator

Marcy E. Rosenkrantz
Director, Library Systems
Digital Library and Information Technologies

Mary Ochs
Head, Collection Development and Preservation, Mann Library
Chair, Publisher Relations Group

Tim Lynch
Head, Information Technology Section, Mann Library
Chair, Technical Design Group

Nancy McGovern
Coordinator, Digital Imaging & Preservation Research

Gregory Lawrence
Government Information Librarian, Mann Library

Bill Kehoe
Programmer-Analyst
Digital Library and Information Technologies (D-LIT)

Advisory Committee

Anne R. Kenney
Assistant University Librarian
Instruction and Learning, Research, and Information Services
Director of Programs
Council on Library and Information Resources
Chair, Steering Committee

Janet McCue
Director, Mann Library

H. Thomas Hickerson
Associate University Librarian for Information Technologies and Special Collections

Mariana Wolfner
Professor of Molecular Biology and Genetics

Appendix: Project Harvest USAIN Survey Fall 2001

This is the Cornell University Library Project Harvest Web-based questionnaire described in our recent email. As we stated in the message, your confidential responses will be very important to our project. At any time, you can withdraw from the survey by closing this Web session. The data collected will be tabulated and shared with the survey participants in early September. Thank you for your cooperation.

1. Your present location is: Please select a 2-letter state abbreviation:

- AK
- AL
- AR
- AZ
- CA

2. Your principle occupation is (choose one):

a. Director/administrator
b. Librarian
c. Educator
d. Student
e. Extension specialist
f. Researcher
g. Information or communication specialist
h. Other (please specify):

3. Please estimate how many e-journals your organization provides access to:

a. 50 or less
b. 51-100
c. 101-250
d. 251-500
e. More than 500

4. How valuable are e-journals to your library in serving its user community (choose one):

a. Minimal value compared to print versions
b. Useful for dual access, cannot substitute for print versions
c. Useful for sole access, can substitute for print versions

5. Would you cancel print journal versions if a reliable archiving solution was available for e-journals?

a. Yes
b. No
c. Other (please specify):

6. Would you have greater trust in the long-term integrity of e-journals if preservation responsibilities were held by:

a. Publisher only
b. Library only
c. A third-party other than a library or publisher
d. Some combination of the above

7. E-journal content can be archived in a very simple form, at low cost, or in more complex forms, at progressively higher cost. Rank the options listed below (1 most worthwhile — 4 least worthwhile):

___ Preserving the basic content only (text and illustrations) in a form that will look different from the original

___ Preserving the original appearance of the journal (the basic content plus the "look and feel" of the publication)

___ Preserving the full functionality of the journal as it appeared initially (basic content plus the "look and feel" plus features like automated reference linking)

___ Preserving the basic content and making it available in a continually updated form consistent with the appearance and functionality of recent issues of the journal

8. Which electronic format would you consider best for archiving e-journals?

a. PDF
b. HTML
c. XML
d. Other (please specify)

9. Have you observed significant information losses in e-journals or other digital resources?

1. Yes
2. No
3. Not sure

10. PLEASE ENTER ANY COMMENTS YOU HAVE FOR US: