YEA: The Yale Electronic Archive
One Year of Progress
Report on the Digital Preservation Planning Project

A collaboration between
Yale University Library and Elsevier Science

Funded by the Andrew W. Mellon Foundation

New Haven, CT
February 2002

Librarianship is a curious profession in which we select materials we don't know will be wanted, which we can only imperfectly assess, against criteria which cannot be precisely defined, for people we've usually never met and if anything important happens as a result we shall probably never know, often because the user doesn't realize it himself. — Charlton quoted by Revill in presentation by Peter Brophy, Manchester Metropolitan University, 4th Northumbria Conference, August 2001

Table of Contents

The Project Team

(Project members began in January 2001 and are continuing with the project unless otherwise indicated)

Yale University Library

Scott Bennett, University Librarian (Principal Investigator, January – July 2001)
Paul Conway, Director of Preservation (Project Manager, January – June 2001)
David Gewirtz, Senior Systems Manager, Yale ITS (Project Technical Director)
Fred Martz, Director of Library Systems (Project Technical Advisor)
Ann Okerson, Associate University Librarian (Co-Principal Investigator, January – July 2001; Principal Investigator July 2001 –)
Kimberly Parker, Electronic Publishing & Collections Librarian (Metadata Investigator)
Richard Szary, Director of University Manuscripts & Archives (Investigator for Archival Uses)

Additional advice and support from:

Matthew Beacom, Catalog Librarian for Networked Information, Yale Library
Jean-Claude Guédon, Professor of Comparative Literature and History of Sciences, Université de Montreal
James Shetler, Asst. Head of Acquisitions, Yale Library
Rodney Stenlake, Esq., Independent Legal Consultant
Stephen Yearl, Digital Systems Archivist, Yale Library

Elsevier Science

Karen Hunter, Senior Vice President for Strategy
Geoffrey Adams, Director of IT Solutions
Emeka Akaezuwa, Associate Director, Information Technology Implementation

Additional advice and support from:

Haroon Chohan, Elsevier Science, IT Consultant
Paul Mostert, Senior Production IT Manager, Hybrid/Local Solutions, ScienceDirect, Elsevier Amsterdam

Acknowledgments

Particular thanks go to the following:

Scott Bennett, for his thoughtful and elegant framing of the issues from the outset to the midpoint of the Project and for always keeping our team on target and on time. From the point of his retirement as of July 31, 2001, we have sincerely missed his dedication and contributions to digital preservation, in which he believed passionately.

The Andrew W. Mellon Foundation, for a tangible demonstration of faith, both in the scholarly community's ability to tackle and begin to solve the vital issues associated with long-term digital preservation, and in the ability of the Yale Library to be one of the players helping to find solutions. Particularly we thank Don Waters of the Foundation for his deep commitment to electronic archiving and preservation and for his help to us.

Our team counterparts at Elsevier Science, for proving to be true partners, giving unstintingly their commitment, time, and thoughtfulness to our joint effort. They have shared fully information, data, and technology expertise. We have learned that our two entities, working together, are much stronger than the sum of our parts.

Yale Information Technology Services, for donating far more of David Gewirtz's time than we had any right to expect and for offering their enthusiastic support and advice.

Other Mellon planning projects and their staffs, for giving us help and insights along the way.

Finally, I personally wish to thank each of our team members, because everyone, out of commitment to the long-term mission of libraries, excitement about digital information technologies and their potential, the thrill of learning, and genuine respect for each other's contributions, did more than their share and contributed to a strong planning phase.

Ann Okerson, Principal Investigator

Executive Summary

Networked information technology offers immense new possibilities in collecting, managing, and accessing unimaginable quantities of information. But the media in which such information lives are remarkably more ephemeral and fragile than even traditional print media. For such information to have a place in scientific and academic discourse, there must be assurance of long-term preservation of data in a form that can be accessed by users with the kind of assurance they now bring to print materials preserved in libraries. The Yale-Elsevier planning project undertook to study the challenges and opportunities for such preservation posed by a large collection of commercially published scientific journals.

Despite the natural interdependence between libraries and publishers, skepticism remains in both communities about the potential for successful library/publisher collaborations, especially in the electronic archiving arena. The e-archiving planning effort between the Yale University Library and Elsevier Science funded by the Andrew W. Mellon Foundation has resulted in substantial gains in bridging the traditional divide between these two groups and has paved the way for continuing collaboration. The goals of our effort were to understand better the scope and scale of digital journal preservation and to reach a position in which it was possible to identify practical next steps in some degree of detail and with a high level of confidence. We believe we have achieved these goals.

From the outset, the Yale-Elsevier team recognized and respected the important and fundamental differences in our respective missions. Any successful and robust e-archive must be built on an infrastructure created specifically to respond to preservation needs and that can only be done with a clear understanding of those missions. Managing library preservation responsibilities for electronic content while protecting a publisher's commercial interests is thus no small task. We have begun with a mutually-beneficial learning process. Work during the Mellon planning year gave us a better understanding of the commercial life cycle of electronic journals, of the ways in which journal production will impact the success of an e-archive, and of the motives that each party brings to the process and the benefits that each party expects.

From the start, the exploration was based on the premise of separating content from functionality. Embedded here is the belief that users of the e-archive are not bench scientists for whom ease of use of the most current scientific information is critical. We envision potential users of the e-archive to be focused primarily on the content. They must be confident that it remains true to what was published and is not influenced/affected by changes in technology that undoubtedly affect functionality. Minimally acceptable standards of access can be defined without mirroring all features of a publisher's evolving interface.

Our determinations include the following:

- Migration of data offers a more realistic strategy than emulation of obsolete systems;
- Preservation metadata differs from that required for production systems and adds real value to the content;
- No preservation program is an island, hence success will depend on adherence to broadly accepted standards and best practices;
- A reasonable preservation process is one that identifies clearly the "trigger events" that would require consultation of the archive and plans accordingly.

We have made effective use of the information learned by library peers in their efforts and in turn have shared the results of our work. Ultimately, the future of electronic archives depends fundamentally on a network of cooperating archives, built on well-conceived and broadly-adopted standards.

Nevertheless, the relationship between publisher and archiver is fundamental. We have begun work on a model license that draws on Yale's extensive experience in developing and modeling successful license agreements. Such an agreement will shape the publisher/archive relationship in ways that control costs, increase effectiveness, and give the archive a place in the economic and intellectual life cycle of the journals preserved.

The Yale-Elsevier team has demonstrated that working collaboratively, we can now begin to build a small prototype archive using emerging standards and available software. This prototype has the potential to become the cornerstone of an e-journal archive environment that provides full backup, preservation, refreshing, and migration functions. We have demonstrated that the prototype — offering content from many or all of the more than 1,200 Elsevier Science journals — can and will function reliably. We are guardedly optimistic about the economic prospects for such archives, but can only test that optimism against a large-scale prototype.

For this archive to become a reality, we must play a continuing lead role in the development and application of standards; help shape and support the notion of a network of cooperating archives; explore further the potential archival uses; and understand and model the economic and sustainability implications.

The following report provides in some detail the results of the Mellon planning year. We believe it demonstrates the deep commitment of the Yale University Library and Elsevier Science to the success of this kind of collaboration, the urgency with which such investigations must be pursued, and the value that can be found in thus assuring the responsible preservation of critical scientific discourse.

Part I: Background, Approaches, Assumptions

Challenges for Long-Term Electronic Archiving

The Big Issues

The tension that underlies digital preservation issues is the fundamental human tension between mutability and immortality. The ancient Platonic philosophers thought that divine nature was unchanging and immortal and human nature was changeable and mortal. They were right about the human nature at least.

Human products in the normal course of affairs share the fates of their makers. If they are not eradicated, they are changed in ways that run beyond the imagination of their makers. Stewart Brand's *How Buildings Learn*[1] has lessons for all those who imagine they are involved in the preservation of cultural artifacts, not just people concerned with buildings.

But a very limited class of things has managed a special kind of fate. The invention of writing and the development of cultural practices associated with it has created a unique kind of survival. It is rarely the case that the original artifact of writing itself survives any substantial length of time, and when it does, that artifact itself is rarely the object of much active reading. Most of us have read the "Declaration of Independence" but few do so while standing in front of the signed original in the National Archives.

Written texts have emerged as man-made artifacts with a peculiar kind of near-immortality. Copied and recopied, transforming physically from one generation to the next, they remain still somehow the same, functionally identical to what has gone before. A modern edition of Plato is utterly unlike in every physical dimension the thing that Plato wrote and yet functions for most readers as a sufficient surrogate for the original artifact. For more modern authors, where the distance between original artifact and present surrogate is shorter, the functional utility of the latter is even greater.

That extraordinary cultural fact creates extraordinary expectations. The idea that when we move to a next generation of technologies we will be able to carry forward the expectations and practices of the last generation blithely and effortlessly is probably widely shared — and deeply misleading. The shift from organic forms of information storage (from papyrus to animal skin to paper) back to inorganic ones (returning, in silicon, to the same material on which the ancients carved words meant to last forever and in fact, lasting mainly only a few decades or centuries) is part of a larger shift of cultural practices that puts the long-term survival of the text newly at risk.

Some of the principal factors that put our expectations in a digital age at risk include:

1. The ephemeral nature of the specific digital materials we use — ephemeral both in that storage materials (e.g., disks, tapes) are fragile and of unknown but surely very limited lifespan, and in that storage technologies (e.g., computers, operating

systems, software) change rapidly and thus create reading environments hostile to older materials.

2. The dependence of the reader on technologies in order to view content. It is impossible to use digital materials without hardware and software systems compatible with the task. All the software that a traditional book requires can be pre-loaded into a human brain (e.g., familiarity with format and structural conventions, knowledge of languages and scripts) and the brain and eyes have the ability to compensate routinely for errors in format and presentation (e.g., typographical errors). The combined effect of those facts makes it impossible for digital materials to survive usefully without continuing human attention and modification. A digital text cannot be left unattended for a few hundred years, or even a few years, and still be readable.

3. The print medium is relatively standard among disciplines and even countries. A physicist in Finland and a poet in Portugal expect their cultural materials to be stored in media that are essentially interchangeable. A digital environment allows for multiple kinds of digital objects and encourages different groups to pursue different goals and standards, thus multiplying the kinds of objects (and kinds of hardware and software supporting different kinds of things) that various disciplines can produce and expect to be preserved.

4. Rapidity of change is a feature of digital information technology. That rapidity means that any steps contemplated to seek stability and permanence are themselves at risk of obsolescing before they can be properly adopted.

5. The intellectual property regimes under which we operate encourage privatization of various kinds, including restricted access to information as well as the creation of proprietary systems designed to encrypt and hide information from unauthorized users until that information no longer has commercial value, at which point the owner of the property may forget and neglect it.

6. The quantity of works created in digital form threatens to overwhelm our traditional practices for management.

7. The aggregation of factors so far outlined threatens to impose costs for management that at this moment we cannot estimate with even an order of magnitude accuracy. Thus we do not have a way of knowing just how much we *can* hope to do to achieve our goals and where some tradeoffs may be required.

8. Finally, the ephemeral nature of the media of recording and transmission imposes a particular sense of urgency on our considerations. Time is not on our side.

Against all of these entropic tendencies lies the powerful force of expectation. Our deepest cultural practices and expectations have to such an extent naturalized ideas of preservation, permanence, and broad accessibility that even the most resistant software manufacturers and anxious owners of intellectual property will normally respond

positively, at least in principle, to concerns of preservation. That is a great advantage and the foundation on which this project is built.

Social and Organizational Challenges

The reader is the elusive center of our attention, but not many readers attend conferences on digital preservation. We find ourselves instead working among authors, publishers, and libraries, with occasional intervention from benevolent and interested foundations and other institutional agencies. The motives and goals of these different players are roughly aligned, but subtly different.

Authors: Authors require in the first instance that their work be made widely known and, for scholarly and scientific information, made known in a form that carries with it the authorization of something like peer review. Raw mass dissemination is not sufficient. Authors require in the second instance that what they write be available to interested users for the useful life of the information. This period varies from discipline to discipline. Authors require in the third instance that their work be accessible for a very long time. This goal is the hardest to achieve and the least effectively pursued, but it nevertheless gives the author community a real interest in preservation. However, the first and second areas of concern are more vital and will lead authors to submit work for publication through channels that offer those services first. In practice, this means that the second motive — desire to see material remain in use for some substantial period of time — is the strongest authorial intent on which a project such as ours can draw. It will drive authors towards reputable and long-standing publishing operations and away from the local, the ephemeral, and the purely experimental.

It should be noted that at a recent digital preservation meeting,[2] certain authors affirmed volubly their right to be forgotten, i.e., at the least to have their content not included in digital archives or to be able to remove it from those archives. Curiously, even as we worry about creating long-term, formal electronic archives that are guaranteed to last, we note that the electronic environment shows quite a multiplier effect: once a work is available on the Web, chances are it can be relatively easily copied or downloaded and shared with correspondents or lists (a matter that worries some rights owners immensely even though those re-sends are rarely of more than a single bibliographical object such as an article). This means that once an author has published something, the chances of his or her right to be forgotten are as slim as they were in days of modern printing — or even slimmer. Think, if you will, whether "special collections," as we have defined them in the print/fixed-format era, have a meaning in the digital environment and, if so, what might that meaning be? The focus may be not on the materials collected so much as on the expertise and commitment to collect and continue to collect materials at particular centers of excellence, especially where the ongoing task of collection is complex, exacting, difficult, and/or particularly unremunerative.

Publishers: Publishers require in the first instance that they recruit and retain authors who will produce material of reliable and widely-recognized value. Hence, publishers and authors share a strong common interest in peer review and similar systems whose

functioning is quite independent of any question of survival and preservation. Publishers are as well — perhaps even more — motivated by their paying customers, e.g., the libraries in this case, who can influence the strategic direction of the publisher no less directly. Consequently, publishers require in the second instance that the material they publish be of continuing value in a specific way. That is, publishers of the type we are concerned with here publish serials and depend for their revenue and the intellectual continuity of their operations on a continuous flow of material of comparable kind and quality. The demand for such material is itself conditioned on that continuity, and the availability of previous issues is itself a mark and instrument of that continuity. The publisher, therefore, understands that readers want the latest material but also material that is not only the latest; the scientific journal is thus crucially different from a newspaper. Publishers also have more insubstantial motivations about continuing reputation and may have themselves long histories of distinguished work. But as with authors, it is the second of the motives outlined here — where self-interest intersects the interest of others — that will most reliably encourage publishers to participate in preservation strategies.

Libraries: Libraries require in the first instance that users find in their collections the materials most urgently needed for their professional and academic work. This need leads them to pay high prices for current information, even when they could get a better price by waiting (e.g., buying new hard cover books rather than waiting for soft cover or remainder prices, or continuing to purchase electronic newspaper subscriptions rather than depend on content aggregators such as Lexis-Nexis, which they may also be purchasing). They require in the second instance that they build collections that usefully reflect the state and quality of scientific, scholarly, and cultural discourse in their chosen areas of coverage.

Research library collections may be built largely independently of day-to-day use, but libraries expect certain kinds of use over a long period of time. Where information may be published to meet a particular information need, these libraries will retain that information in order to meet their other mission of collecting a cultural heritage for future generations. Yale has been pursuing that project for 300 years. With traditional media, libraries, museums, archives, and other cultural institutions have pursued that project in various independent and cooperative ways. It is reasonable to expect that such cooperation among traditional players and new players will only continue and grow more powerful when the objects of preservation are digital.

All of the above tantalizing complications of long-term electronic archiving drew us into this planning project and continued to be themes throughout the year-long planning process.

Background for the Planning Project

Yale Library as a Player

The Yale Library has, for close to three decades, been cognizant of and aggressive about providing access to the numerous emerging forms of scholarly and popular publications appearing in "new" electronic formats, in particular those delivered via the Internet and increasingly through a Web interface. Initially, indexing and abstracting services and other "reference" type works found their way into electronic formats and rapidly displaced their former print instantiations. By 1996, full-text content from serious and substantive players such as Academic Press (AP) and JSTOR were entering the marketplace, and libraries — and their readers — became quickly captivated by the utility, effectiveness, efficiency, and convenience of academic content freed from its traditional fixed formats. By the summer of 2000, Yale Library, like most of its peer institutions, was spending over $1 million annually on these new publication forms, offering several hundred reference databases and thousands of full-text electronic journals to its readers. Expenditures for electronic content and access paid to outside providers last academic year alone totaled nearly $1.8 million. In addition, we spend increasing sums both on the creation of digital content and on the tools to manage digital content internally — e.g., a growing digitized image collection, a fledgling collection of university electronic records, digital finding aids for analog materials in our collections, and our online library management system, which includes but is not limited to the public access catalog.

The electronic resources that Yale makes available to readers, both on its own and in conjunction with its partners in the NorthEast Research Libraries consortium (NERL), quickly become heavily used and wildly popular — and increasingly duplicative, in terms of effort and price, with a number of print reference works, journals, and books. The Yale Library, with its significant resources, 300 years of collections, and long-term commitment to acquiring and preserving collections not only for its own but also for a global body of readers, has acted cautiously and prudently in retaining print texts not only for immediate use but also for long-time ownership and access. It has treated its growing, largely licensed (and thus un-owned) electronic collections, moreover, as a boon for reader productivity enhancement and convenience, even to the point of acquiring materials that seem to duplicate content but offer different functionality.

Nonetheless, it is clear that the Library (and this is true of all libraries) cannot endlessly continue on such a dual pathway, for several compelling reasons: 1) the growing level of print and electronic duplication is very costly in staff resources and to collections budgets; 2) increasingly, readers, at least in certain fields such as sciences, technology, and medicine (STM), as well as certain social sciences, strongly prefer the electronic format and the use of print in those areas is rapidly diminishing; and 3) traditional library stacks are costly to build or renovate. All libraries face what NERL members have dubbed the "Carol Fleishauer"[3] problem: "We have to subscribe to the electronic resources for many good reasons, but we cannot drop print — where we might wish to —

because we do not trust the permanence of electronic media." The challenge for us all, then, is how to solve the problem that Carol so pointedly articulated at a NERL meeting.

An Opportunity to Learn and Plan

When the Andrew W. Mellon Foundation first began to signal its keen interest in long-term preservation of digital resources, the Yale Library had already acquired, loaded on site, and migrated certain selected databases and full-text files for its readers. The Library had also explored, between 1997 and 1999, the possibility of serving as a local repository site for either all of Elsevier Science's more than 1,100 full-text e-journal titles or at least for the 300 or so that had been identified by NERL members as the most important for their programs. Yale information technology (IT) staff experimented with a year's worth of Elsevier electronic journals, using the Science Server software for loading and providing functionality to those journals. In the end, the fuller functionality of Science Direct (Elsevier's commercial Web product) proved more attractive for immediate access, particularly as interlinking capabilities between publishers were developed and were more easily exploitable from the commercial site than through a local load. Nonetheless, Yale's positive experience in working with Elsevier content led our staff enthusiastically to consider Elsevier as a possible e-journal archive planning partner.

When we were invited to consider applying for a planning grant (summer 2000) we contacted Karen Hunter, Elsevier Science's Senior Vice President for Strategy, who signaled a keen personal and corporate interest in long-term electronic archiving solutions. Additional attractions for Yale in working with Elsevier related to the huge amount of important content, particularly in STM, that this publisher provides. It also seemed to us that the strong commercial motivation of a for-profit publisher made such a partnership more important and interesting, at least initially, than one with a not-for-profit publisher. That is, a for-profit entity might not feel so much of a long-term or eternal commitment to its content as would a learned society. We have learned in the course of this planning project that Elsevier has its own serious commitment in this area. With Ms. Hunter and other senior Elsevier staff, a self-identified Yale Library team began to formulate e-journal questions and opportunities, and Elsevier officers strongly supported our submission of the Fall 2000 application to Mellon.

Throughout the project, regular meetings between Yale and Elsevier team members have been held, and key topics have been identified and pursued therein, as well as by phone, e-mail, through the work of small sub-groups, through site visits by team members and visitors to our establishments. The principal lines of inquiry have been in the following areas: 1) "trigger" events; 2) a beginning exploration of the fascinating area of "archival uses," i.e., how real-life users might use the kind of archive we were proposing to develop; 3) contractual and licensing issues; 4) metadata identification and analysis, particularly through comparison and cross-mapping with datasets recommended by such players as the British Library and OCLC/RLG; and 5) technical issues, beginning with an examination of Elsevier's production and workflow processes, and with particular emphasis after summer of 2001 on building a small prototype archive based on the OAIS (Open Archival Information Systems) model,[4] focusing on the archival store

component. Interlaced have been the economic and sustainability questions that are crucial to all electronic preservation projects.

An Interesting Mid-Year Development: Acquisition of Academic Press (AP)

In July 2001, the acquisition by Elsevier Science of a number of Harcourt properties and their component journal imprints was announced. This real-life business event presented interesting and significant challenges, not only to Elsevier Science but also prospectively to the Yale Electronic Archive (YEA). At that time, Elsevier had not fully formulated its organizational or technical plans for the new imprints, which included not only AP (whose IDEAL service represented one of the earliest commercially-licensed groups of important scientific journals, introduced to the marketplace in 1995), but also Harcourt, Saunders, Mosby, and Churchill Livingstone, i.e., primarily medical titles which are to be integrated into a new Elsevier division called Health Science. Elsevier staff believe that the imprints of the acquired titles will survive; it is not clear whether the IDEAL electronic platform will continue or whether (more likely) it will be integrated into Science Direct production processes.

Sizing

As a result of this acquisition, Elsevier's total number of journal titles rose from around 1,100-1,200, to about 1,500-1,600, i.e., by nearly 50 percent. In 2002, the combined publications will add collectively 240,000 new articles to Elsevier's electronic offerings; the size of the Elsevier e-archive will become 1.7 million articles in number. Additionally, Elsevier, like many other scholarly and scientific publishers, is pursuing a program of retrospective digitization of its journals, back to Volume One, Number One. The backfiles up to and including the years 1800-1995 (estimated at four to six million entries) will require 4.5 terabytes of storage. The years up to and including 1995-2000 require .75 terabytes. Production output for 2001 is estimated to be at 150-200 gigabytes using the new ScienceDirect OnSite (SDOS) 3.0 format, SGML, and graphics. The addition of Harcourt titles adds another 1.25 terabytes of storage requirements, for a grand total of nearly 7 terabytes of storage being required for all Elsevier Science content back to Volume One, Number One.

Although our e-archiving project is not yet working with non-journal electronic formats, it is worth noting that just as with Elsevier Science's publishing output, the new AP acquisitions include serials such as the Mosby yearbooks, various *Advances in*…, and certain major reference works. The Yale team believes that this non-journal output is worthy of attention and also poses significant challenges.

After the acquisition, Ken Metzner, AP's Director of Electronic Publishing, was invited to attend project meetings; scheduling permitted him to participate only to a limited extent. Because of the relative newness of this additional content, the YEA team participants were unable to determine the specific impacts of the acquisition for our electronic archiving pursuits. Details will become clearer during 2002. What is clear — and has been all along, though not so dramatically — is that content of publishers is fluid, at least

around the edges. Titles come and go; rights are gained and lost, bought and sold. Any archive needs carefully to consider its desiderata with regard to such normally occurring business changes and how contractual obligations must be expressed to accommodate them.

Approaches and Assumptions

In January 2001, the Mellon Foundation approved a one-year planning grant for the Yale Library in partnership with Elsevier Science. The two organizations issued a press release and named members to the planning team. The work proceeded according to certain assumptions, some of which were identified very early in the process and others that evolved as we began to pursue various lines of inquiry. While it is not always easy to distinguish, one year into the planning process, assumptions from findings, we have made an effort to do so and to detail our *a priori* assumptions here:

1. The digital archive being planned will meet long-term needs. The team adopted a definition of "long" as one hundred or more years. One hundred years, while rather less than the life of well cared for library print collections, is significantly more than any period of time that digital media have so far lasted or needed to last. Users of an electronic archive must be secure in the knowledge that the content they find in that archive is the content the author created and the publisher published, the same kind of assurance they generally have with print collections.

2. Accordingly, and given the rapid changes in technologies of all sorts, the archive has the responsibility to migrate content through numerous generations of hardware and software. Content can be defined as the data and discourse about the data that authors submit to the publisher. Content includes, as well, other discourse-related values added by the publisher's editorial process, such as revisions prompted by peer review or copy editing or editorial content such as letters, reviews, journal information and the like. Specifically, content might comprise articles and their abstracts, footnotes, and references, as well as supplementary materials such as datasets, audio, visual and other enhancements. For practical purposes, of course, the content consists of the finished product made available to readers.

3. The archive will not compete with the publisher's presentation of or functionality for the same content nor with the publisher's revenue stream. Functionality is defined as a set of further value-adding activities that do not have a major impact on the reader's ability to read content but may help the reader to locate, interact with, and understand the content. The YEA project will not concern itself with reproducing or preserving the different instances of publisher's or provider's functionality, because this functionality is very mutable and appears differently based on who delivers the content (publisher, vendor, aggregator, and so on). That is, an increasing number of journals, books, and other databases are provided through more than one source and thus multiple interfaces are already available, in many instances, for the same works.

4. Should the archive be seen as potentially competitive, the publisher would certainly have no incentive (and the authors might not have any incentive either) to cooperate with the archive by contributing to it, particularly content formatted for easy ingestion into the archive.

5. That said, some immediate uses of the archive seem desirable. We imagined that if/when the archive — or some portion of its offerings, such as metadata — could be conceived of as having rather different and possibly more limited uses than the primary, extensive uses the publisher's commercial service provides, the archive could be deployed early in its existence for those uses.

6. Once each set of data is loaded at regular, frequent intervals, the archive remains an entity separate from and independent of the publisher's production system and store. The publisher's data and the archive's data become, as it were, "fraternal twins" going their separate behavioral ways.

7. Such environmental separation enables the archive's content to be managed and rendered separately and to be migrated separately — and flexibly — over long periods of time, unencumbered by the formatting and production processes that the publisher needs to deploy for its own print and electronic dissemination.

8. The archive, accordingly, commits to preserving the author's content and does not make an effort to reproduce or preserve the publisher's presentation of the content, providing, at this time, basic functionality and content with no visible loss. The YEA is committed at this point in time only to a minimum "no frills" standard of presentation of content.

9. Where extensive functionality is required, the YEA project assumes that functionality will be created — perhaps by the archive or by a separate contractor — when the time comes, and that the functionality will be created in standards applicable to that (present or future) time.

10. The archive does not, in general, create electronic content that does not exist in the publisher's electronic offering of the same content, even though such content may exist in the printed version of the journals. The archive is not intended to mimic the printed version. (For example, if the print journal includes "letters to the editor" and the e-journal version does not include these, the e-archive will not create them.

11. The archive will likely need to create metadata or other elements to define or describe the publisher's electronic content if the publisher has not provided these data during its production processes. However, it is desirable for the archive to create as few of these electronic elements as possible. The most accurate and efficient (and cost-effective) archive will be created if the publisher creates those data. This in turn indicates strongly the need both for industry-wide electronic preservation standards and close partnerships between archives (such as libraries) and publishers.

12. The archive will work with the publisher to facilitate at-source creation of all electronic elements. The archive will work with other similar publishers and archives to develop, as quickly as possible, standards for such elements, in order to deliver consistent and complete archive ingestion packages.

13. The archive will develop a system to ingest content regularly and frequently. In this way, it will record information identical to that generated by the authors and publishers. Any adjustments to content after ingestion into the archive will be added and identified as such.

14. At Yale, the e-journals archival system and site will be part of a larger digital store and infrastructure comprising and integrating numerous other digital items, many of which are already on site, such as internally (to the University and Library) digitized content, images, University records, born-digital acquisitions that are purchased rather than leased, finding aids, preservation re-formatting, the online catalog, and others. In this way, efficiencies and synergies can be advanced and exploited.

15. The archive will regularly and frequently "test" content, preferably both with automated systems that verify accuracy and completeness and with real users seeking content from the archive.

16. YEA team members assume that the archive may be searched by outside intelligent agents or harvesters or "bots," and that it must be constructed in a way that both facilitates such searching and at the same time respects the rights agreements that have been made with the copyright owners.

17. "Triggers" for ingestion are frequent and immediate; "triggers" for use of the archive are a different matter and will comply with rules developed between the publisher and the archive. These triggers would be identified during the course of the planning project.

18. The archive will be constructed to comply with emerging standards such as the OAIS model, certain metadata recommendations if possible, XML, and the like. Standards that enable data portability are key to YEA development.

19. The archive will be developed using, wherever possible, software tools as well as concepts being created in other institutions.

Notes About Archival Approaches and the Absolute Necessity for Standards in E-Archival Development

The assumptions listed above speak to four major activities of the YEA, which are 1) preservation planning, 2) administration of the archive, 3) access to the archive, and 4) ingestion of content. Like other Mellon research projects, the YEA defines activities associated with these processes within the context of the OAIS reference model of a

digital archive. *A fortiori*, the model also states that implementations will vary depending upon the needs of an archival community.

Research conducted during the planning year has identified four different approaches to preservation: emulation, migration, hard copy, and computer museums. These different approaches should not, however, be viewed as mutually exclusive. They can be used in conjunction with each other. Additionally, as we are not pursuing the present study as simply an academic exercise but rather, as a very practical investigation of what it will take to build, operate, and maintain a fully functioning production-mode digital archive, we cannot possibly discount the financial implications of the different approaches and the impact of standards on a choice of approach.

In choosing our approach, we quickly discounted both hard copy and computer museums. Hard copy decays over time and multimedia objects cannot be printed. Computer museums were also discounted as impractical. To function as an archive, computer museum equipment would need to be operational, not simply a collection of static exhibits. In turn, operation would lead to inevitable wear and tear on the equipment, with the consequential need for maintenance and repair work. When considering the repair of "antique" computer equipment, one has to ask about the source of spare parts — do all of them have to be hand-made at enormous expense? Even if money were available for such expensive work, does the "antique" equipment come with adequate diagnostic and testing equipment, wiring diagrams, component specifications, and the like, which would make the "museum" choice technically feasible in the first place? Even if the antique equipment were to receive only the most minimal usage, the silicon chips would deteriorate over time. In such a museum environment, one would question whether we would be in danger of losing our focus, ending up as a living history-of-computers museum rather than an archive of digital materials.

Rejecting hardcopy and museum options left the team with two very different approaches to the storage of content in an OAIS archive. One approach to content preservation is to store objects based upon emerging standards such as XML and then migrate them to new formats as new paradigms emerge. The other approach, advanced by and associated with one of its chief proponents, Marc Rothenberg, is to preserve content through emulation. Both approaches have potential and value, but probably to different subcultures in the archival community. A goal of standards is to preserve the essential meaning or argument contained in the digital object. Archives that are charged with the responsibility of preserving text-based objects such as e-journals are likely to adopt a migratory approach. Archives that need to preserve an exact replication or clone of the digital objects may choose, in the future, to deploy emulation as an archival approach. Both approaches are rooted in the use of standards. Contrary to an argument advanced in a research paper by Rothenberg,[5] YEA team members maintain that standards, despite their flaws, represent an essential component of any coherent preservation strategy adopted.

That said, Rothenberg criticized the use of migration as an approach to digital longevity. Rothenberg does make an insightful and enterprising case for the practice of emulation as the "true" answer to the digital longevity problem. The "central idea of the approach, " he

writes, "is to enable the emulation of obsolete systems on future, unknown systems, so that a digital document's original software can be run in the future despite being obsolete." Rothenberg avers that only by preservation of a digital object's context — or, simply stated, an object's original hardware and software environment — can the object's originality (look, feel, and meaning) be protected and preserved from technological decay and software dependency.

The foundation of this approach rests on hardware emulation, which is a common practice in the field of data processing. Rothenberg logically argues that once a hardware system is emulated, all else just naturally follows. The operating system designed to run on the hardware works and software application(s) that were written for the operating system also work. Consequently, the digital object behaves and interacts with the software as originally designed.

However, emulation cannot escape standards. Processors and peripherals are designed with the use of standards. If the manufacturer of a piece of hardware did not adhere 100 percent to the standard, then the emulation will reflect that imperfection or flaw. Consequently, there is never a true solution, as suggested by Rothenberg, that a generalized specification for an emulator of a hardware platform can be constructed. In the data processing trenches, system programmers are well acquainted with the imperfections and problems of emulation. For example, the IBM operating system MVS never ran without problems under IBM's VM operating system. It was a good emulation but it was not perfect. Another major problem with emulation in a practical sense is its financial implications. The specification, development, and testing of an emulator require large amounts of very sophisticated and expensive resources.

At this stage, the YEA team believes the most productive line of research is a migratory approach based upon standards. Standards development must, therefore, feature front and center in the next phase of e-journal archiving activities. If one listens closely to academic discourse, the most seductive adverb of all is one not found in a dictionary; it is spelled "just" and pronounced "jist" and is heard repeatedly in optimistic and transparent schemes for making the world a better place. If scientists would "jist" insist on contributing to publishing venues with the appropriate high-minded standards of broad access, we would all be better off. If users would "jist" insist on using open source operating systems like Linux, we would all be better off. If libraries would "jist" spend more money on acquisitions, we would all be better off.

Many of those propositions are undoubtedly true, but the adverb is their Achilles' heel. In each case the "jist" masks the crucial point of difficulty, the sticking point to movement. To identify those sticking points reliably is the first step to progress in any realistic plan for action. In some cases, the plan for action itself is arguably a good one, but building the consensus and the commonality is the difficulty; in other cases, the plan of action is fatally flawed because the "jist" masks not merely a difficulty but an impossibility.

It would be a comparatively easy thing to design, for any given journal and any given publisher, a reliable system of digital information architecture and a plan for preservation

that would be absolutely bulletproof — as long as the other players in the system would "jist" accept the self-evident virtue of the system proposed. Unfortunately, the acceptance of self-evident virtue is a practice far less widely emulated than one could wish.

It is fundamental to the intention of a project such as the YEA that the product — the preserved artifact — be as independent of mischance and the need for special supervising providence as possible. That means that, like it or not, YEA and all other seriously aspiring archives must work in an environment of hardware, software, and information architecture that is as collaboratively developed and as broadly supported as possible, as open and inviting to other participants as possible, and as likely to have a clear migration forward into the future as possible.

The lesson is simple: standards mean durability. Adhering to commonly and widely recognized data standards will create records in a form that lends itself to adaptation as technologies change. Best of all is to identify standards that are in the course of emerging, i.e., that appear powerful at the present moment and are likely to have a strong future in front of them. Identifying those standards has an element of risk about it if we choose the version that has no future, but at the moment some of the choices seem fairly clear.

Standards point not only to the future but also to the present in another way. The well-chosen standard positions itself at a crossroads, from which multiple paths of data transformation radiate. The right standards are the ones that allow transformation into as many forms as the present and foreseeable user could wish. Thus PDF is a less desirable, though widely used, standard because it does not convert into structured text. The XML suite of technology standards is most desirable because it is portable, extensible, and transformative: it can generate everything from ASCII to HTML to PDF and beyond.

Plan of Work

The Project Manager chart describes the planning project's working efforts during the year, and it highlights certain key events: http://www.diglib.org/preserve/2bplan.pdf.

Part II: Lines of Inquiry

Trigger Events

Makers of an archive need to be very explicit about one question: what is the archive for? The correct answer to that question is not a large idealistic answer about assuring the future of science and culture but a practical one: when and how and for what purpose will this archive be put to use? Any ongoing daily source needs to be backed up reliably, somewhere away from the risks of the live server, and *that* backup copy becomes the *de facto* archive and the basis for serious preservation activities.

Types of Archives

The team discovered during the course of its explorations that there is no single type of archive. While it is true that all or most digital archives might share a common mission, i.e., the provision of permanent access to content, as we noted in our original proposal to the Mellon Foundation, "This simple truth grows immensely complicated when one acknowledges that such access is also the basis of the publishers' business and that, in the digital arena (unlike the print arena), the archival agent owns nothing that it may preserve and cannot control the terms on which access to preserved information is provided."

In beginning to think about triggers, business models, and sustainability, the project team modeled three kinds of archival agents. The first two types of archives include a *de facto* archival agent, defined as a library or consortium having a current license to load all of a publisher's journals locally, or a self-designated archival agent. Both of these types are commercial transactions, even though they do not conduct their business in the same ways or necessarily to meet the same missions. The third type of archive is a publisher-archival agent partnership and the focus of our investigation. Whether this type can now be brought into existence turns on the business viability of an archive that is not heavily accessed. Project participants varied in their views about whether an archive with an as yet uncertain mission can be created and sustained over time and whether, if created, an individual library such as Yale or a wide-reaching library enterprise like OCLC would be the more likely archival partner.

Accessing the Archive

So when does one access the archive? Or does one ever access it? If the archive is never to be accessed (until, say, material passes into the public domain, which currently in the United States is seventy years plus the lifetime of the author or rights holder), then the incentives for building it diminish greatly, or at least the cost per use becomes infinite. There is talk these days of "dark" archives, that is, collections of data intended for no use but only for preservation in the abstract. Such a "dark" archive concept is at the least risky and in the end possibly absurd.

Planning for access to the e-archive requires two elements. The less clearly defined at the present is the technical manner of opening and reading the archive, for this will depend on the state of technology at the point of need. The more clearly defined, however, will be what we have chosen to call "trigger" events. In developing an archival arrangement with a publisher or other rights holder, it will be necessary for the archive to specify the circumstances in which 1) the move to the archive will be authorized, which is much easier to agree to than the point at which 2) users may access the archive's content. The publisher or rights holder will naturally discourage too early or too easy authorization, for then the archive would begin to attract traffic that should go by rights to the commercial source. Many rights holders will also naturally resist thinking about the eventuality in which they are involuntarily removed from the scene by corporate transformation or other misadventure, but it is precisely such circumstances that need to be most carefully defined.

Project participants worked and thought hard to identify conditions that could prompt a transfer of access responsibilities from the publisher to the archival agent. These conditions would be the key factors on which a business plan for a digital archive would turn. The investigation began by trying to identify events that would trigger such a transfer, but it concluded that most such events led back to questions about the marketplace for and the life cycle of electronic information that were as yet impossible to answer. Team members agreed that too little is known about the relatively young business of electronic publishing to enable us now to identify definitively situations in which it would be reasonable for publishers to transfer access responsibility to an archival agent.

Possible Trigger Events

That said, some of the possible trigger events identified during numerous discussions by the project team were:

Long-term physical damage to the primary source. Note that we have not imagined the e-journal archive to serve as a temporary emergency service. We expect formal publishers to make provision for such access. Nevertheless, in the case of cataclysmic event, the publisher could have an agreement with the archive that would allow the publisher to recopy material for ongoing use.

Loss of access or abdication of responsibility for access by the rights holder or his/her successor, or no successor for the rights holder is identified. In other words, the content of the archive could be made widely available by the archive if the content is no longer commercially available from the publisher or future owner of that content. We should note that at this point in time, we were not easily able to imagine a situation in which the owner or successor would not make provision precisely because in the event of a sale or bankruptcy, content is a primary transactional asset. But that is not to say that such situations will not occur or that the new owner might not choose to deal with the archive as, in some way, the distributor of the previous owner's content.

Lapse of a specified period of time. That is, it could be negotiated in advance that the archive would become the primary source after a negotiated period or "moving wall," of the sort that JSTOR has introduced into the e-journal world's common parlance. It may be that the "free science" movement embodied in PubMed Central or Public Library of Science might set new norms in which scientific content is made widely available from any and all sources after a period of time, to be chosen by the rights owner. This is a variant on the "moving wall" model.

On-site visitors. Elsevier, our partner in this planning venture, has agreed that at the least, its content could be made available to any onsite visitors at the archive site, and possibly to other institutions licensing the content from the publisher. Another possibility is provision of access to institutions that have previously licensed the content. This latter option goes directly to development of financial and sustainability models that will be key in Phase II.

Archival Uses. Elsevier is very interested in continuing to explore the notion of so-called "archival uses" which represent uses very different to uses made by current subscribers in support of today's active science. Elsevier has stated that if we can identify such "archival uses," it might be willing to consider opening the archive to those. Some such uses might be studies in the history, sociology, or culture of sciences, for example. This thread in our planning processes has motivated the YEA team to devote some time to early exploration of archival uses with a view to expanding and deepening such exploration in Phase II.

Metadata Uses. In the course of preservation activity it could be imagined that new metadata elements and structures would be created that would turn out to have use beyond the archive. Appropriate uses of such data would need to be negotiated with the original rights holder.

Some Economic Considerations

Economic considerations are key to developing systems of digital archives. Accordingly, in our proposal, the Yale Library expressed its intention better "to understand the ordinary commercial life cycle of scientific journal archives..." In that proposal, our list of additional important questions included concerns about costs of creating and sustaining the archive, as well as sources of ongoing revenues to support the archive. While the issues of sustainability lurked in our thinking throughout the project, we determined relatively early on that the time was not right substantively to address these matters because we had as yet insufficient data and skill to make any but the very broadest of generalizations. But, that lack of hard data did not stop the group from discussing and returning frequently to economic matters.

Neither were the views of various individuals and organizations of definitive help to us. For example, the best study about e-archiving known to us attempted to analyze costs, but the information is somewhat dated.[6] A large school of thought affirms that e-archives and even e-journal archives will be immensely expensive to develop and maintain, perhaps impossibly so. Some of the arguments include:

Huge Costs. Formal publishers' e-journal titles, i.e., those presented in fairly "standard" formats, will be very costly to archive because even those publishers do not provide clean, consistent, fully tagged data. Accordingly, the e-archive will have to perform significant repair and enhancement, particularly in the ingestion process; e.g., the creation of the Submission Information Package (SIP) will be particularly expensive. Furthermore, this reasoning goes, as the size, variety, and complexity of the content increases, associated costs will rise, as they will whenever formats need to be migrated and as storage size increases.

The universe of e-journals — which includes a great volume as well as diversity of subjects and formats, including Web sites, newsletters, dynamic publications, e-zines, and scholarly journals, and includes a huge variety of possible technical formats — will surely be difficult and costly to archive when one considers that universe as a whole.

Information Will Be Free. On the other hand, a great deal of today's "popular" scientific literature, promulgated by working scientists themselves, argues that electronic archiving is very cheap indeed. Proponents of this optimistic line of argument reason that colleges, universities, research laboratories, and the like already support the most costly piece of the action: that electronic infrastructure comprises computers, internal networks, and fast links to the external world, and institutions are obligated in any case aggressively to maintain their investments and frequently to update them. That being the case, the reasoning is that willing authors can put high quality material "out there," leaving it for search engines and harvesters to find. In such arguments, the value-adding services heretofore provided by editors, reviewers, publishers, and libraries are doomed to obsolescence and are withering away even as this report is being written.

Our guess is that the "truth" will be found to lie in between those two polarities, but of course that guess is a little glib and perhaps even more unfounded than the above arguments.

Even though during the planning year we were unable to make economic issues a topic of focused inquiry, we have begun to develop specific and detailed costs for building the YEA for e-journals in preparation for the next granting phase, and those calculations are starting to provide us with a sense of scale for such an operation. In addition, throughout the year, team members articulated certain general views about the economics of e-journal archives, which we share here below.

Five Cost Life-Cycle Stages of an e-Journal Archive

The task of archiving electronic journals may be divided into five parts: the difficult part (infrastructure development and startup), the easier part (maintenance), the sometimes tricky part (collaborations and standards), the messy part (comprehensiveness), and the part where it becomes difficult again (new technologies, migration).

1. **The difficult part (development and startup).** Initial electronic archiving efforts involve such activities as establishing the data architecture, verifying a prototype, validating the assumptions, and testing the adequacy of the degree of detail of realization. The magnitude and complexity of the issues and the detail involved in e-journal archiving are considerable. That said, it does not lie beyond the scope of human imagination, and the big lesson we have learned in this planning year is that it is indeed possible to get one's arms around the problem, and that several different projects have discovered more or less the same thing in the same time period. In fact, Yale Library is already involved in other types of archiving projects related to several other digital initiatives. The greatest difficulties do not lie in having to invent a new technology, nor do they lie in coping with immense magnitudes. Rather, they reside in resolving a large, but not unimaginably large, set of problems in an adequate degree of detail to cope with a broad range of possibilities.

2. **The easier part (ongoing maintenance and problem resolution).** Where we are encouraged is in believing that once the first structure-building steps have been

taken, the active operationalization and maintenance of an e-journal archiving project, in partnership with one or more well-resourced and cooperative publishers, can become relatively straightforward, particularly as standards develop to which all parties can adhere. There will be costs, but after start-up many of these will be increasingly marginal costs to the act of publishing the electronic journal in the first place. For new data being created going forward, attaching appropriate metadata and conforming to agreed standards will require up-front investment of time and attention, especially retrofitting the first years of journals to standards newly enacted, but once that is done, the ongoing tasks will become more transparent. In theory, the hosting of the archive could be part and parcel of the operational side of the publishing, and the servers and staff involved in that case would most likely be the same people involved in the actual publication. Alternately, as we imagine it, the long-term archiving piece of business will be taken aboard by existing centers distributed among hosting universities with similar synergies of costs.

3. **The tricky part (collaboration and standards).** Because different people and organizations in different settings have been working on electronic preservation issues for the last few years, there may already be appreciable numbers of similar but nonidentical sets of solutions coming to life. Working around the world to build sufficient communities of interest and standards to allow genuinely interoperable archives and real standards will take a great deal of "social work." Every archive will continue to devote some percentage of its operation to external collaborations driven by the desire to optimize functional interoperability.

4. **The messy part (comprehensiveness)**. There will be a fair number of journals that either choose not to cooperate or are financially or organizationally ill-equipped to cooperate in a venture of the scope imagined. It will be in the interest of the library and user communities generally to identify those under-resourced or recalcitrant organizations and find the means — financial, organizational, political — to bring as many of them aboard as possible. It may prove to be the case that 90 percent of formal publishers' journals can be brought aboard for a modest price, and the other 10 percent may require as much money or more to come in line with the broader community.

5. **The part where it becomes difficult — and probably very expensive — again (migration).** The solutions we now envision will sustain themselves only as long as the current technical framework holds. When the next technological or conceptual revolution gives people powers of presentation they now lack and that do not allow themselves to be represented by the technical solutions we now envision, then we will require the next revolution in archiving. The good news at that point is that some well-made and well-observed standards and practices today should be able to be carried forward as a subset of whatever superset of practices need to be devised in the future. Elsevier Science has a foretaste of this in its current, very costly migration to XML.

Needless to say, the above overview is somewhat simplified. For example, in our planning year, we were surprised to find just how few of the 1,100+ Elsevier e-journal titles carried complex information objects, compared to what we expected to find. Complex media, data sets, and other electronic-only features exist that have yet to find their place as regular or dominant players in e-journals, and creating ways to deal with these types of digital information — let alone standard ways — will be costly, as are all initial structural activities (see #1 above).

Cost-Effective Collaboration and Organization for e-Archiving

That said, it appears that willing collaborators have yet a little time both to address and to solve the hefty problems of presenting and archiving complex digital information objects. To archive a single e-journal or small set of journals is to do relatively little. But to develop standards that will serve e-preservation well — let alone to facilitate access to the most simple of e-archives that begin to bloom like a hundred flowers — all the players will need to work together. We imagine an aggregation of archiving efforts, whether in physical co-location or at least virtual association and coordination.

But how might such archival universes be organized?

- Archives could be subject-based, arranged by discipline and subdiscipline. Such an arrangement would allow some specialization of features, easier cross-journal searching, and creation of a community of stakeholders.

- Archives could be format-based. This arrangement would probably overlap with subject-based arrangement in many fields, would be easier to operate and manage, but would sacrifice at least some functionality for users — an important consideration, given that archival retrieval is likely to occur in ways that put at least some demand on users to navigate unfamiliar interfaces.

- Archives could be publisher-based. Such an arrangement would offer real conveniences at the very outset, but would need close examination to assure that standards and interoperability are maintained against the natural interest of a given rights holder to cling to prerogatives and privileges.

- Archives could be nationally-based. Australia, Japan, Canada, Sweden, and other nations could reasonably feel that they have a mission to preserve their own scientific and cultural products and not to depend on others.

- Archives could be organized entrepreneurially by hosts. This is probably the weakest model, inasmuch as it would create the least coherence for users and searching.

Each of these alternate universes has its own gravitational force and all will probably come into existence in one form or another. Such multiplicity creates potentially severe problems of scalability and cost. One remedy could be for official archives to operate as

service providers feeding other archives. Hence, a publisher's agreed archive could feed some of its journals to one subject-based archive and others to national archives.

One way to begin to anticipate and plan for this likely multiplicity would be to create a consortium now of interested parties to address the difficult issues such as redundancy, certification, economic models, collection of fees, standards, and so on. No one organization can solve these problems alone, but coordination among problem-solvers now and soon will be very cost-effective in the long run. In OCLC's proposal to create a digital preservation cooperative,[7] and, on a larger scale in the Library of Congress's recent National Digital Information Infrastructure Preservation Program,[8] we may be seeing the emergence of such movements. It may be possible to turn the Mellon planning projects into such an overarching group of groups.

Who Will Pay and How Will They Pay?

No preservation ambitions will be realized without a sustainable economic model. As we have noted above, the costs of archiving are much in dispute and our study will examine those costs in great detail in the next phase. For now, it would appear that the initial costs are high, although manageable, and the ongoing costs, at least for standard publisher's journals, could be relatively predictable and eventually stable over time.

If that is true, then various models for paying for the archiving process suggest themselves. This is an area about which there has been much soft discourse but in which there has been little experience, save perhaps for JSTOR whose staff have given the topic a great deal of thought.

Up-front payment. The most dramatic and simple way to finance the e-journal archives would be the "lifetime annuity model": that is, users (presumably institutional entities, such as libraries, professional societies, governments, or cultural institutions, but some speak of enhanced "page charges" from authors or other variants on current practices) pay for a defined quantum of storage and with that one-time payment comes an eternity of preservation. The up front payment would be invested partly in ongoing archival development and partly in an "endowment" or rainy day fund. The risk in this case is that inadequate funding may lead to future difficulties of operation.

Ongoing archival fees. An "insurance premium" on the other hand could give an ongoing supply of money, adjustable as costs change, and modest at all stages. This reduces the risk to the provider but increases the uncertainty for the beneficiary. The ongoing fee could be a visible part of a subscription fee or a fee for services charged by the archive.

The traditional library model. The library (or museum or archive) picks up the tab and is funded by third-party sources.

Fee for services operation. The archive provides certain services (special metadata, support for specialized archives) in return for payments.

Hybrid. If no single arrangement seems sufficient — as it likely will not — then a hybrid system likely will emerge, perhaps with one set of stakeholders sharing the up-front costs while another enters into agreement to provide ongoing funding for maintenance and potential access.

Much more could be said on the topic of who pays but at the moment most of it would be speculation. The choice of models will influence development of methods for paying fees and the agents who will collect those fees. Before making specific recommendations it will be important for our project to develop a much more specific sense of real costs of the e-archive. We imagine that we might want to develop both cost and charging models in conjunction with other libraries, i.e., prospective users of the archive. In Yale's case the collaborative effort might happen with our local electronic resource licensing consortium NERL.

Contract between the Publisher and the Archive

Publishers and librarians have reluctantly grown accustomed to having licenses that articulate the terms and conditions under which digital publications may be used. These licenses are necessary because in their absence the uses to which digital files could be put would be limited by restrictions (and ambiguities) on reproduction and related uses that are intrinsic within copyright law. Licenses clarify ambiguities and often remove, or at least significantly reduce, limitations while also acknowledging certain restrictions on unlimited access or use.

A licensing agreement between a digital information provider and an archival repository presents several unique challenges not generally faced in the standard licensing agreement context between an information provider and an end-user. Discussed below are several of the issues that must be addressed in any final agreement:

Issues

1. Term and termination. The perpetual nature of the intended agreement, even if "forever," is in fact, a relative rather than an absolute term. One has to think in funereal terms of "perpetual care" and of the minimum length of time required to make an archiving agreement reasonable as to expectations and investments. Some issues that need to be addressed are appropriate length of any such agreement, as well as provisions for termination of the agreement and/or "handing off" the archive to a third party. Underlying concerns of term and termination is the need to ensure that the parties' investments in the archive are sufficiently protected as well as that the materials are sufficiently maintained and supported.

2. Sharing responsibility between the archive and the digital information provider. There are elements of a service level agreement that must be incorporated into the license because the rights and responsibilities are different in an archival agreement than in a normal license. That is, an archive is not the same as a traditional end-user; in many ways the archive is stepping into the shoes of the

digital information provider in order (eventually) to provide access to end-users. The rights and responsibilities of the archive will no doubt vary depending on when the material will become accessible and on whether there are any differentiations between the level and timing of access by end-users. This issue will have an impact on the level of technical and informational support each party is required to provide to end-users and to each other, as well responsibility for content — including the right to withdraw or change information in the archive — and responsibilities concerning protecting against the unauthorized use of the material.

3. Level and timing of access. While all licenses describe who are the authorized users, the parties to an archival agreement must try to anticipate and articulate the circumstances (i.e., "trigger events") under which the contents of the archive can be made available to readers, possibly without restriction. When the information will be transmitted to the archive and, more importantly, how that information is made available to end-users are also critical questions. Several models have been discussed and this may be an issue best addressed in detailed appendices reflecting particular concerns related to individual publications.

4. Costs and fees. The financial terms of the agreement are much different from those of a conventional publisher-user license. Though it is difficult to conceive of one standard or agreed financial model, it is clear that an archival agreement will have a different set of financial considerations from a "normal" license. Arrangements must be made for the recovery of costs for services to end-users, as well as any sharing of costs between the archive and the digital information provider. These costs may include transmission costs, the development of archive and end-user access software, and hardware and other costs involved in preserving and maintaining the data.

5. Submission of the materials to the archive. The issues of format of the deposited work ("submission") take on new considerations as there is a need for more information than typically comes with an online or even locally-held database. Describing the means for initial and subsequent transfers of digital information to the archive requires a balance between providing sufficient detail to ensure all technical requirements for receiving and understanding the material are met, while at the same time providing sufficient flexibility for differing technologies used in storing and accessing the materials throughout the life of the contract. One means of dealing with the submission issues is to provide in the agreement general language concerning the transmission of the materials, with reference to appendices that can contain precise protocols for different materials in different time periods. If detailed appendices are the preferred method for dealing with submission matters, mechanisms must be developed for modifying the specifics during the life of the agreement without triggering a formal renegotiation of the entire contract.

6. Integrity of the archive. The integrity and comprehensiveness of the archive must be considered. The contract must address the question: "If the publisher 'withdraws' a publication, is it also withdrawn from the archive?"

Progress Made

YEA and Elsevier Science have come to basic agreement on what they would be comfortable with as a model license. In some areas alternatives are clearly available and other archival agencies working with other publishers will choose different alternatives. Reaching a general agreement was, however, surprisingly easy as the agreement flowed naturally out of the year-long discussions on what we were trying to accomplish. The current draft license is not supplied in this document because it has a number of "unpolished" areas and some unresolved details, but it could be submitted and discussed upon request.

The team made certain choices with regard to the contractual issues noted above:

1. **Term**. The team opted for an initial ten-year term with subsequent ten-year renewals This provides the library with sufficient assurance that its investments will be protected and assures the publisher that there is a long-term commitment. The team also recognized that circumstances can change and has attempted to provide for what we hope will be an orderly transfer to another archival repository.

2. **Rights and responsibilities.** The agreement includes statements of rights and responsibilities that are quite different from a traditional digital license. The publisher agrees, among other things, to conform to submission standards. The library agrees, among other things, to receive, maintain, and migrate the files over time.

3. **Trigger events**. Discussions of "trigger events" provided some of the most interesting, if also frustrating, aspects of the year. In the end, the only trigger event that all completely agreed upon was that condition under which the digital materials being archived were no longer commercially available either from the original publisher or someone who had acquired them as assets for further utilization. Given that it is quite hard to imagine a circumstance in which journal files of this magnitude would be judged to have no commercial value and would not be commercially offered, does it makes sense to maintain such an archive at all? Will money be invested year after year as a precaution or protection against an event that will never occur? Though the team agreed it is necessary to proceed with long-term electronic archival agreements, clearly serious issues are at stake.

   The team also identified a second side to the trigger question: if the archive were not going to be exposed to wide use by readers, how could the archival agent "exercise" it in order to assure its technical viability? This topic is discussed more fully in the "Trigger Events" section of the report. Briefly here, the team was concerned that a totally dark archive might become technically unusable over

time and wanted to provide agreed upon applications that would make the archive at least "dim" and subject to some level of use, e.g., available to local authorized users. The second, perhaps more important, notion was that there would be archival uses that could be distinguishable from normal journal use. The team tried to identify such uses but so far have not received the feedback from the history of science community (for example) that we would have wished. Therefore, "archival uses" remain more theory than reality, but at the same time they represent a topic we are committed to exploring in the next phase of work. An alternative would be to have the archive serve as a service provider to former subscribers, but this changes the nature of the archive to being a "normal" host which could be a questionable consideration. These issues are not currently reflected in the draft license.

4. **Financial terms** were viewed as neutral at this time, i.e., no money would change hands. In our current thinking, the publisher provides the files without charge and the archival agency accepts the perpetual archiving responsibility without financing from the publisher. Obviously, one could argue that the publisher should be financing some part of this activity. However, in the longer term it is probably more realistic to develop alternative financing arrangements that are independent of the publisher.

5. **Technical provisions**. Early on, the team agreed on the OAIS model for submission and subsequent activities. The license reflects this in terms of the need to define metadata provided by the publisher. The specific metadata elements have not yet been finalized, however. This is also relevant in defining what use can be made by the archive of the metadata. Publishers such as Elsevier that have secondary publishing businesses want to be sure that those businesses are not compromised by an archive distributing abstracts for free, for example. The model license does not yet reflect this point but it is recognized as an issue.

6. **Withdrawal of content**. The current draft license provides for appropriate notices when an item is withdrawn by the publisher. The team has discussed and will likely incorporate into the license the notion that the archive will "sequester" rather than remove a withdrawn item.

The model license is still evolving and not yet ready for signature. However, there are no identified points of contention — only points for further reflection and agreement on wording. All the participants were very much pleased with the team's ability to come to early understandings of licensing issues and to resolve some of these at the planning stage. This success arises out of close working relationships and communications over about a year-and-a-half of cooperative effort.

Archival Uses of Electronic Scientific Journals

As part of its work, the Yale-Elsevier team began to investigate whether and how the uses of an archive of electronic journals would differ significantly from those of the active product distributed by the publisher. This investigation was launched to help determine

what needed to be preserved and maintained in the archive; to inform the design of a discovery, navigation, and presentation mechanism for materials in the archive; and to determine the circumstances under which materials in the archive could be made available for research use without compromising the publisher's commercial interests.

The group reviewed traditional archival theory and practice and began preliminary consultations with historians of science and scholarly communication to understand past and contemporary uses of scientific journal literature. A number of issues became particularly significant in the group's discussions: the selection of documentation of long-term significance, the importance of topological and structural relationships within the content, and the importance of the archive as a guarantor of authenticity.

Selection and Appraisal

The first area in which there might be useful approaches is that of archival appraisal, i.e., the selection of those materials worth the resources needed for their long-term preservation and ongoing access. Archival appraisal considers the continuing need of the creating entity for documentation in order to carry out its mission and functions and to maintain its legal and administrative accountability, as well as other potential uses for the materials. These other uses generally fall into the category of support for historical research, although there may be others such as establishing and proving the existence of personal rights which may also be secondary to the original purpose of the documentation in question.

Archivists also consider the context of the documentation as well as its content in determining long-term significance. In some cases, the significance of the documentation lies in the particular content that is recorded; the presentation of that content is not critical to its usefulness or interpretation. The content of the documentation can be extracted, put into other applications, and made to serve useful purposes even as it is divorced from its original recording technology and form. In other cases, however, the role of documentation as evidence requires that the original form of the document and information about the circumstances under which it was created and used also be preserved in order to establish and maintain its authenticity and usefulness.

With these selection approaches in mind, a number of issues arose in the e-journal archiving context and in the work of the team. The first question was whether it was sufficient for the archive to preserve and provide access to "just" the content of the published material — primarily text and figures — in a standard format, which might or might not be the format in which the publisher distributed the content. Preserving only the content, insofar as that is possible, foregoes the preservation of any functionality that controlled and facilitated the discovery, navigation, and presentation of the materials on the assumption that functionality was of little or no long-term research interest. The decision to preserve content only would eliminate the need to deal with changing display formats, search mechanisms and indices, and linking capabilities.

While the group has adopted this narrow definition of the scope of the archive as a working assumption, such a narrow approach does preclude the study of the diplomatics of these documents — "digital paleography," as one of our advisors termed it. How essential to future researchers' interpretations of the use of these documents is it for them to know what tools contemporary users had available to them, e.g., indices that did not address particular components of the document, thus making them unfindable through the publisher's interface? At the conclusion of the planning period the team had not changed the main focus of its attention on content, but it was sufficiently intrigued by the issues of digital paleography that it will propose that this assumption be investigated more thoroughly in its implementation proposal.

The long-held approach in the archival profession governing how archives are organized, described, and provided to users once they become part of the repository's holdings is deeply informed by the principle of provenance and the rules that flow from it: *respect des fonds* (records of a creator should remain together) and original order (which has significance for the interpretation of records and should be preserved whenever possible). These principles reflect the nature of archival records. They are by-products created by an organizational entity in the course of carrying out its functions. The primary significance of the records is as evidence of those functions and activities. These principles reflect the needs of research for bodies of materials that are as strongly evidential as possible and reflect minimal interaction by custodial agencies other than the creator. The assumption is that solid historical research and interpretation require knowledge of the circumstances under which the materials were created and maintained and not just access to the raw content.

Access to archival materials is often characterized by two factors that take advantage of the provenance approach. Searches are often conducted to document a particular event or issue rather than for a known item; they may also be based on characteristics of the creators rather than on characteristics of the records themselves. Comprehensive and accurate recording of the circumstances of creation, including characteristics of records creators and the relationships among them, are central parts of archival description. The implications for developing an approach to downstream uses of e-journal literature include the potential need of contextual metadata regarding the authors and other circumstances affecting the publication of a given article/issue that are not found in a structured way in the published materials. Information regarding the context in which the article was submitted, reviewed, and edited for publication is important in studies of scholarly communication, especially as to questions of how institutional affiliations might be important in certain lines of inquiry and who had the power to accept or reject submissions.

Some of this information is explicitly disseminated in online products, e.g., in the form of members of an editorial board or descriptions of the purpose and audience of the journal, but it may be presented separately from any particular volume, issue, or article; may reflect only current (and not historic) information; and is rarely structured or encoded in such a way as to facilitate its direct use in scholarly studies. Other information about the context of creation and use that historians of science might find useful is not published;

rather, it is found in the publisher's records of the review process and circulation figures. Capturing and linking of title-level publication information are additional areas of investigation that the team intends to pursue in its implementation proposal.

Preservation of Structural Information

The mass of archival records that repositories select for long-range retention and are responsible for, and the imperative of the principle of provenance to maintain and document the recordkeeping system in which the records were created and lived, combine to foster the archival practice of top-down, hierarchical, and collective description. This type of descriptive practice provides both a way of reflecting the arrangement of the original recordkeeping system and of allowing the archival agency to select for each body of records the level beyond which the costs of description outweigh the benefits of access, and completing its descriptive work just before that point is achieved.

This principle and practice highlight for scientific journals the importance of preserving the relationship among the materials that the publisher was distributing, especially the need to link articles that the publisher presented as a "volume," "special issue," or some other sort of chronological or topical grouping. These relationships represent another form of contextual information important to the study of scholarly communications, in terms of which articles were released simultaneously or in some other relationship to each other. While the team recognized the need to be aware of new forms of publishing that would not necessarily follow the traditional patterns adopted by the hard-copy print world, it asserted that those structures do need to be saved as long as they are used.

With respect to other methods of navigating among digitally presented articles, such as linking to articles cited, the team found that many of these capabilities existed not as part of the content, but as added functionality that might be managed by processes external to the content or to the publisher's product (e.g. CrossRef). The team felt that these capabilities should be preserved as part of the archive, necessitating the need to maintain an enduring naming scheme for unambiguous identification of particular pieces. The plan for the implementation project will include a closer look at the requirements for supporting important navigational capabilities.

Guaranteeing Authenticity

Finally, the authenticity of any document that purports to be evidence rests in some part on a chain of custody that ensures that the document was created as described and that it has not been altered from its original form or content. Once an archival agency takes charge of documentation it is obligated to keep explicit records documenting the circumstances of its transfer or acquisition and any subsequent uses of it. Records are rarely removed, either for use or retrospective retention by an office, but when this is necessary the circumstances of that action need to be documented and available. This assumption, along with the unique nature and intrinsic value of the materials, leads to the circumstance of secure reading rooms for archival materials and all of the security

paraphernalia associated with them, as well as to detailed recordkeeping of use and work performed on the records.

The assumption that the archival agency is responsible for preserving the "authentic" version of documentation suggests that transfer of content to the official archival agency should take place as soon as the publisher disseminates such content, and that once placed into the archive content will not be modified in any way. This includes instances of typographical errors, the release of inaccurate (and potentially dangerous) information, or the publication of materials not meeting professional standards for review, citation, and similar issues. Instead, the archive should maintain a system of errata and appropriate flagging and sequestering of such materials that were released and later corrected or withdrawn, ensuring that the record of what was distributed to the scholarly community, however flawed, would be preserved.

Issues related to authenticity also suggest that one circumstance under which transferred content could be released, even while the publisher retains a business interest in it, is when questions are raised as to the authenticity of content still available under normal business arrangements. Longer-term safeguards will need to be in place within the archival repository to ensure the authenticity of the content.

Other issues relating to the nature and mission of an archival repository appear elsewhere in this report, especially in the discussion of trigger events. The issues discussed in this section, however, are especially germane to the question of how anticipated use of preserved electronic journals should inform the selection of materials. The Yale-Elsevier team has found many archival use topics central to the definition and purpose of an archive for electronic journals and plans to pursue them more completely in the implementation project.

The Metadata Inquiry

The Role of Metadata in an e-Archive

It is impossible to create a system designed to authenticate, preserve, and make available electronic journals for an extended period of time without addressing the topic of metadata. "Metadata" is a term that has been used so often in different contexts that it has become somewhat imprecise in meaning. Therefore, it is probably wise to begin a discussion of metadata for an archival system by narrowing the array of possible connotations. In the context of this investigation, metadata makes possible certain key functions:

- Metadata permits search and extraction of content from an archival entity in unique ways (descriptive metadata). Metadata does this by describing the materials (in our case journals and articles) in full bibliographic detail.

- Metadata also permits the management of the content for the archive (administrative metadata) by describing in detail the technical aspects of the

> ingested content (format, relevant transformations, etc.), the way content was ingested into the archive, and activities that have since taken place within the archive, thereby affecting the ingested item.

Taken together, both types of metadata facilitate the preservation of the content for the future (preservation metadata). Preservation ensures the retrievability of protected materials, their authentication, and their content.

Using metadata to describe the characteristics of an archived item is important for a number of reasons. With care, metadata can highlight the sensitivity to technological obsolescence of content under the care of an archival agency (i.e., items of a complex technical nature that are more susceptible to small changes in formats or browsers.) Metadata can also prevent contractual conflicts by pinpointing issues related to an archived item's governance while under the care of an archive; e.g., "the archive has permission to copy this item for a subscriber but not for a nonsubscriber." Finally, metadata can permit the archival agency to examine the requirements of the item during its life cycle within the archive; e.g., "this object has been migrated four times since it was deposited and it is now difficult to find a browser for its current format."[9]

The Open Archival Information System (OAIS) model to which the YEA project has chosen to conform refers to metadata as preservation description information (PDI). There are four types of PDI within OAIS: 1) reference information, 2) context information, 3) provenance information, and 4) fixity information. Not all of these forms of PDI need be present in the Submission Information Package (SIP) ingested by the archive, but they all must be a part of the Archival Information Package (AIP) stored in the archive. This implies that some of these PDI elements are created during ingestion or input by the archive.

Reference Information refers to standards used to define identifiers of the content. While YEA uses reference information and supplies this context in appendices to our metadata element set, we do not refer to it as metadata. Context Information documents the relationships of the content to its environment. For YEA, this is part of the descriptive metadata. Provenance Information documents the history of the content including its storage, handling, and migration. Fixity Information documents the authentication mechanisms and provides authentication keys to ensure that the content object has not been altered in an undocumented manner. Both Provenance and Fixity are part of administrative metadata for YEA.

Given the focus YEA has chosen to place on a preservation model that serves as an archive as well as a guarantor for the content placed in its care, authenticity was an issue of importance for the group to explore. In its early investigations, the team was much struck by the detailed analysis of the InterPARES project on the subject of authenticity. While some of the InterPARES work is highly specific to records and manuscripts — and thus irrelevant to the journal archiving on which YEA is focusing — some general principles remain the same. It is important to record as much detail as possible about the original object brought under the care of the archive in order both to prove that a

migrated or "refreshed" item is still the derivative of the original and to permit an analysis to be conducted in the future about when and how specific types of recorded information have changed or are being reinvented, or where totally new forms are emerging.[10]

Finally, as YEA has examined the issue of metadata for a system designed to authenticate, preserve, and make available electronic journals for an extended length of time, we have tried to keep in mind that metadata will not just be static; rather, metadata will be interacted with, often by individuals who are seeking knowledge. To this end, we acknowledge the four issues identified by the International Federation of Library Associations and Institutions (IFLA) in the report on functional requirements for bibliographic records: metadata exist because individuals want to find, identify, select, or obtain informational materials.[11]

The Metadata Analysis

YEA began its analysis of needed metadata for a preservation archive of electronic journals by conducting a review of extant literature and projects. In this process the team discovered and closely explored a number of models and schemes. The first document — and the one we returned to most strongly in the end — described the OAIS model, although OAIS provides only a general framework and leaves the details to be defined by implementers.[12] We also examined the Making of America's testbed project white paper[13] and determined it was compatible with OAIS. Next, we examined the 15 January 2001 RLG/OCLC Preservation Metadata Review document[14] and determined that while all of the major projects described (CEDARS, NEDLIB, PANDORA) were compliant with the OAIS structure, none of them had the level of detail, particularly in contextual information, that we believed necessary for a long-term electronic journal archive. We also explored the InterPARES project (mentioned above) and found there a level of detail in contextual information that we had not seen delineated in the RLG/OCLC review of the other projects.

At the same time, the library and publisher participants in the project were exploring the extant metadata sets used by Elsevier Science to transport descriptions of their journal materials for their own document handling systems and customer interfaces. In addition to their EFFECT standard (see section describing Elsevier Science's Technical Systems and Processes), we also examined portions of the more detailed Elsevier Science "Full Length Article DTD 4.2.0."[15] Due to the solid pre-existing work by Elsevier Science in this area and the thorough documentation of the metadata elements that Elsevier Science is already using, we were able to proceed directly to an analysis of the extant Elsevier metadata to determine what additional information might need to be created or recorded during production for and by YEA.

About halfway through the project year, the team made connections with British Library staff who were themselves just completing a metadata element set definition project and who generously shared with the team their draft version. While the British Library draft document was more expansive in scope than the needs of the YEA project (i.e., the

British Library document covers manuscripts, films, and many other items beyond the scope of any e-journal focus), the metadata elements defined therein and the level of detail in each area of coverage were almost precisely on target for what the e-archiving team sought to create. Thus, with the kind consent of contacts at the British Library, the team began working with the draft, stripping away unneeded elements, and inserting some missing items.

In the fall of 2001, the YEA team committed to creating a working prototype or proof-of-concept which demonstrated it would indeed be possible to ingest data supplied by Elsevier Science into a minimalistic environment conducive to archival maintenance. The prototype-building activity briefly diverted the metadata focus from assembling a full set of needed elements for the archival system to defining a very minimal set of elements for use in the prototype. The technical explorations of the prototype eventually led us to simply use the metadata supplied by Elsevier and the prototype metadata element set was never used. The one remaining activity associated with metadata performed for the prototype was to map the Elsevier EFFECT metadata to Dublin Core so that it could be exposed for harvesting.

Once the prototype subset element set was identified, YEA returned to the question of a complete metadata element set for a working archive. As the British Library draft document was examined, reviewed, and assessed, many decisions were made to include or exclude elements, particularly descriptive metadata elements. These decisions were informed in part by the recurring theme of whether the presence of such an item of information would assist individuals performing inquiries of the archive. The questions related to uses of scholarly journal materials for archival explorations are dealt with more fully elsewhere in this report.

The full metadata element set was completed by YEA as a recommended set of metadata to be used in a future full archive construction. It is important to reiterate that our approach to producing this set of metadata was inclusive. In creating an archival architecture it is not enough to delineate the descriptive metadata that must be acquired from the publisher or created by the archive while leaving out the administrative metadata elements that permit the archive to function in its preserving role. Neither is it sufficient to focus on the administrative metadata aspects that are unique to an archive while setting aside the descriptive metadata elements, i.e., assuming they are sufficiently defined by other standards. Preservation metadata are the conflation of the two types of metadata and, in fact, both types of metadata work jointly to ensure the preservation of and continuing access to the materials under the care of the archive.

One other fact may be of interest to those reviewing the description of metadata elements for the YEA: where possible, we used external standards and lists as materials upon which the archive would depend. For example, we refer to the DCMI-Type Vocabulary[16] as the reference list of the element called "resource type."

We certainly do not expect that the element set created by YEA will proceed into implementation in a future full archive construction without any further changes. It will

undoubtedly be influenced by work done by other groups such as the *E-Journal Archive DTD Feasibility Study*[17] prepared for the Harvard University Library e-journal archiving project. However, we now have a reference by which to assess whether proposed inclusions, exclusions, or modifications fit the structure we imagine an archive will need properly to preserve electronic journals.

Metadata in Phase II

In the next phase of the e-archiving project, the YEA desires further to define and refine metadata needed for a system designed to authenticate, preserve, and make available electronic journals for an extended period of time. We will need to connect with others working informally or formally to create a standard or standards for preservation metadata. As noted above, further investigations may influence a revision to the initial metadata set defined during the planning phase. Additionally, we intend to rework the element set into an XML schema for Open Archives Initiative (OAI) manifestation and harvesting. With our small prototype, we have demonstrated that OAI harvesting can occur from the simple Dublin Core metadata set to which we mapped the Elsevier EFFECT elements. However, OAI interaction can occur much more richly if a fuller dataset is in use, and we intend to accomplish this schema transformation to enable fuller interaction with the archive as it develops.[18]

As the next phase moves forward, another avenue of exploration will be to assess and examine our metadata element choices in a live environment. We are most particularly interested in testing those elements included or excluded on the basis of assumptions made regarding the likelihood of archival inquiries targeting specific elements for exploration. Such choices can only be validated over time and with the interaction of individuals conducting authentic research into the history of their fields. Finally, we look forward to testing our element choices for administrative metadata under the stress of daily archive administration and maintenance. Only in such a live environment can an archive be truly confirmed as a functioning entity for preserving the materials under its care.

Elsevier Science's Technical Systems and Processes

Introduction

Elsevier Science is a major producer of scholarly communication and scientific journals that are distributed globally. The headquarters for production, along with the electronic warehouse, is located in Amsterdam, Netherlands. There the company maintains two office buildings and deploys several hundred staff to organize, produce, and distribute its content. The production of electronic scholarly information is a highly complex process that occurs in a distributed geographical environment involving many businesses beyond Elsevier Science. Changes to the manufacturing process can take years to percolate through the entire chain of assembly and are considered significant business risks. Consequently, Elsevier is moved to make changes to production only when compelling market demands exist. For example, the Computer Aided Production (CAP) workflow is

now under modification because Science Direct, an internal customer of Elsevier Science, is experiencing market pressure to bring published items to its customers in a shorter time than ever before.

The History

Prior to the creation of the Electronic Warehouse (EW) in 1995, Elsevier Science had no standard processes to create and distribute journals or content. The production of journals was based upon a loose confederation of many smaller publishing houses owned by Elsevier Science. Content was produced using methods that were extant when Elsevier acquired a given publisher. Consequently, prior to the creation of the EW there was no uniformity in the structure or style of content marketed under the name of Elsevier Science. Each publishing house set its own standards for creation and distribution. The lack of a central infrastructure for creating and distributing content also served as an impediment to the rapid distribution of scholarly communication to the market.

With the creation of networks in the early 1990s the perception of time delay amplified. Scientists began to use the network themselves to share communications with one another instantly. The scholarly community would no longer accept long delays between the submission of manuscripts to a publisher and their appearance in paper journals. Scientists and publishers realized that reporting research in electronic format could significantly close the time gap between publication and distribution of content to the scholarly community. The origin of Elsevier Science's Electronic Warehouse is rooted in this realization. Elsevier Science's early solution to the problem was to support a research project known as The University Licensing Program commonly referred to as TULIP.[19]

The TULIP Project (1992-1996) grew out of a series of Coalition for Networked Information (CNI) meetings in which Elsevier Science, a CNI member, invited academic libraries to partner with it to consider how best to develop online delivery capabilities for scientific journals. The purpose of the project was to discuss the need to build large-scale systems and infrastructures to support the production and rapid delivery of such journals over a network to the scholarly community. Given a critical mass of interest from the university communities, Elsevier Science justified a large investment that would create a manufacturing function for converting paper journals into an electronic format for network distribution. This process became known as the PRECAP method for the creation of an electronic version of a journal. The creation of this conversion function served as the foundation for the present day EW. Near the end of the TULIP project plans for an EW were adopted by Elsevier Science in 1995 and built by the end of 1996. By 1997 the EW could produce over one thousand journals using a standard means of production.

The creation and success of the EW in producing and distributing journals was a very significant accomplishment for Elsevier Science because 1) many individual publishers had to be converted one by one, 2) standards for production were evolving from 1994 through 2000, and 3) suppliers who created content for the producers needed to be

continuously trained and retooled to adhere to the evolving standards. At the same time, these suppliers met their obligation to produce content, on time, for Elsevier Science.

Current Workflow

Elsevier Science maintains four production sites based in the United Kingdom (Oxford and Exeter), Ireland (Shannon), the United States (New York), and the Netherlands (Amsterdam). Each site provides content to the EW where this content is stored as an S300 dataset. The contents of each dataset represent an entire issue of a particular journal. The storage system at the EW originally used vanilla IBM technology, i.e., ADSTAR Distributed Storage Manager (ADSM), to create tape backup datasets of content stored on magnetic and optical storage. Access to the data was based only upon the file name of the S300 dataset. As of Summer 2001, the old hierarchical storage system was replaced by an all-magnetic disk-based system providing more flexibility and enabling faster throughput and production times.

The CAP Workflow

The following is a concise description and discussion of the Computer Aided Production (CAP) workflow. An item is accepted for publication by means of a peer review process. After peer review the item enters the CAP workflow via the Login Function in which a publication item identifier (PII) is assigned to the content. This is a tag that the EW uses to track the item through the production process, and it also serves as a piece of metadata used for the long-term storage of the item. Since this identifier is unique it could also be used as a digital object identifier for an information package in an OAIS archive. In addition to assigning the PII, the login process also obtains other metadata about the author and item such as the first author's name, address, e-mail address, and number of pages, tables, and figures in the item. This and other similar metadata are entered into a Production Tracking System (PTS) that is maintained by the Production Control system.

The item is then sent electronically to a supplier (Elsevier has sixteen suppliers, distributed on a worldwide basis). There the item undergoes media conversion, file structuring, copy editing, and typesetting. The output of this processing is a first generation (no corrections) SGML markup of the item, a PDF file, and artwork for the item. These units of work are then sent to the author for corrections. The author makes the necessary corrections, then sends the item to Production Control where information in the PTS system is updated. Thereafter, Production Control sends the item to an Issues Manager. Any problems found in the content are worked out between the author and the Issues Manager. If there are no problems, the supplier sends the content directly to Production Control.

The Issues Manager then passes the corrections on to the supplier and begins to compile the issue. This involves making decisions about proofs, cover pages, advertising, and building of indexes. On average, an Issues Manager is responsible for five to ten journals or about fifteen thousand pages a year. Once content is received, the supplier then creates a second-generation SGML and PDF file and new artwork, if necessary. This cycle is

repeated until the complete issue is assembled. Once the issue is fully assembled the Issues Manager directs the supplier to create a distribution dataset called S300 which contains the entire issue. The supplier sends this file to the EW where the file serves as input for the creation of distribution datasets for customers such as Science Direct. At EW this dataset is added to an ADSM-based storage system that serves as a depository — not an archive — for all electronic data produced for the EW. The S300 dataset is also sent to a printer where a paper version of the issue is created and then distributed to customers. The paper version of the journal is also stored in a warehouse. Most printing occurs in the Netherlands and the United Kingdom.

The current issue-based workflow has two serious problems. The first is that production does not produce content for distribution in a timely fashion for customers like Science Direct, and the second is that issue-based processing generates high and low periods of work for suppliers. A steady stream of work passing through the manufacturing process would be more efficient for the suppliers and would result in a more timely delivery of content to Elsevier's customers such as Science Direct. The driving force behind a need for change, as mentioned above, is not EW but rather, Science Direct as an internal customer of the EW. The resolution to these workflow problems is to change the fundamental unit of work for production from an issue to an article, something Elsevier recognizes and is currently working toward.

The new article-based e-workflow being developed by Science Direct will streamline interactions between authors, producers, and suppliers. At a management level automation of key functions will yield the following efficiencies: 1) in the e-workflow model, Web sites will be created to automate the electronic submission of articles to an editorial office and to establish an electronic peer review system, and 2) the peer review system will interface with a more automated login and tracking system maintained by the EW.

The new Production Tracking System can then be used by the EW, suppliers, and customers to manage the production and distribution processes more efficiently. Functionally, the EW would also produce two additional intermediary datasets called S100 and S200. These datasets could be sent to the EW for distribution to customers at the time of creation by the supplier and before an S300 dataset was sent to the EW. For example, the physics community, which uses letter journals, would directly benefit by this change in production. Under the e-workflow model, the supplier could immediately upon creation send an S100 dataset that contained a first generation version of the letter or item (i.e., no author corrections) directly to the EW for distribution to a Science Direct Web site. In addition, Science Direct would also be able to distribute content at the article level in the form of an S200 dataset that contained second generation or correct SGML and PDF data. This content would be sent to a Web site before an S300 dataset, representing the entire issue that was sent to the EW by the supplier. It is interesting to note that the EW does not save intermediary datasets once an S300 dataset is created. Pilot projects have been launched to test the e-workflow model.

Finally, it should be noted that as the use of the EW developed and evolved over time, it became apparent — for operational and customer support reasons — that some additional support systems would be needed. For example, one of these systems facilitates Elsevier's ability to support customers in auditing the completeness of their collections. Another tracks the history of publications that Elsevier distributes.

The Standards

In the early 1990s Elsevier Science recognized that production and delivery of electronic content could best be facilitated by conversion of documents to an SGML format. SGML is a tool that enables the rendering of a document to be separated from the content structure of a document. This division is achieved through the use of a document type definition (DTD) and a style sheet. The DTD is a tool by which the structure of a document can be defined through the use of mark-up tags. In addition, the DTD defines a grammar or set of rules that constrain how these tags can be used to mark up a document. A style sheet defines how the content should be rendered, i.e., character sets, fonts, and type styles. Together, these two tools make documents portable across different computer systems and more easily manipulated by database applications. In addition, the separation of content from rendering is also critical to the long-term preservation of electronic scholarly information. That said, the evolution of production and distribution of content by Elsevier Science or the EW has been tightly coupled to 1) the development of a universal DTD for their publications, 2) the successful adoption of a DTD by EW suppliers, and 3) the emergence of the Portable Document Format known as PDF. On average it took two years for all suppliers (at one time greater than two hundred) to integrate a new DTD into production. As inferred from Table I below, by the time one DTD was fully implemented by all suppliers another version of the DTD was being released.

Table I — DTD Chronology of Development

DTD Version	Date	Note
FLA 1.1.0	April 1994	
FLA 2.1.1	May 1995	
FLA 3.0.0	November 1995	No production
FLA 4.0.0		No production
Index DTD 1.0.0		No production
Glossary DTD 1.0.0		No production
FLA 4.1.0	November 1997	Full SGML
FLA 4.2.0	February 2000	Full SGML Perfected
FLA 4.3.0	July 2001	
FLA 5.0	To be announced	XML and MathML

Two types of standards control production exist at the EW, one for the production of content and the other for the distribution of content to customers. On the production side, the standards used are known as PRECAP, CAP, and DTD. For distribution to customers such as Science Direct, the standards are known as Exchange Format for Electronic Components and Text (EFFECT), Electronic Subscriptions Service (EES), and Science Direct On Site (SDOS).

PRECAP, born in 1995, is an acronym for PRE-Computer Aided Production; CAP, born in 1997, is an acronym for Computer Aided Production. The principal differences between the two modes of production are that 1) PRECAP is paper-based and CAP is electronically based, and 2) CAP production produces higher quality content then PRECAP production. In the PRECAP method electronic journals are created from disassembled paper journals that are scanned and processed.

The PRECAP standard was developed for the TULIP project and is still used to produce content today. At first, PRECAP content was distributed for the Elsevier Science EES (1995-1997) using the EFFECT dataset format, a standard that also grew out of the TULIP project. In fact, the standard was first known as the TULIP Technical Specification Version 2.2. Like its predecessor, EFFECT provided a means by which output from PRECAP and CAP processing could be bundled and delivered to customers. The specification defined a map or standard by which component data (e.g., TIFF, ASCII, and PDF) generated by these processes could be accessed by an application or loaded into a database for end-user use. Since its introduction in 1995, the EFFECT standard has gone through several revisions to meet the changing needs of EW customers.[20]

The data components for PRECAP production can consist of page image files, raw ASCII files, SGML files for bibliographic information, and postscript files encapsulated PDF format. Data components for CAP production are only different in that CAP contains no TIFF page images and CAP can produce full SGML instances of an editorial item such as a full-length article. Paper pages are scanned at 300 dots per inch (dpi) to produce TIFF 5.0 (Tag Image File Format) images. The page image has a white background and black characters. Pages are also compressed using the Fax Group 4 standard that can reduce the size of a page image by about 8 percent. The raw text files are generated via Optical Character Recognition (OCR) from the TIFF files. No editing is performed on these files and only characters in the ASCII range 32-126 are present in the content. These raw files are used to create indexes for applications, not for end-user purposes. SGML citation files are created using the full-length article DTD version 3.0 that took nearly three years to develop. DTD version 1.1.0 created in 1994 and version 2.1.1 created in 1995 were considered experimental and were not used to produce content with PRECAP production. Sometime in 1996 or early 1997, PDF files replaced TIFF image files.

The development of the PRECAP standard and its evolution to CAP can be traced in datasets the EW distributed to its customers either as EES or later as SDOS. Between 1995 and 1997, the EW distributed three different versions of datasets to the Elsevier Electronic Subscription Service. These versions are known as EES 1.0, EES 1.1, and EES

1.2. In EES 1.0, datasets contained only TIFF, ASCII, and SGML files. In EES 1.1, PDF files replaced TIFF images. However, in EES 1.2 datasets (around 1997) a new type of PDF file was introduced. This PDF file is called a "true PDF" and is created not from paper but rather from output from electronic type setting. This change marked the birth of the CAP production method. It is important to note that both CAP and PRECAP data can be contained in an EES 1.2 distribution dataset and SDOS datasets. In December of 1998 EES was officially renamed SDOS. From about 1997 to the present, three versions of SDOS datasets (SDOS 2.0, SDOS 2.1, and SDOS 3.0) have been marketed. Like EES, SDOS datasets contain the same component data. Differences in the versions consist of the type of SGML content packaged in the dataset. All versions contain SGML bibliographic data but in version 2.1 tail or article reference data is being delivered in SGML format. It also is important to note that by 1997, using DTD 4.1, EW suppliers produced full-length article SGML. However, this content was not offered to the EW customers from 1997 through 2000. In February of 2000, DTD 4.2 became the production DTD. However, it was only recently, in spring of 2001, that SDOS 3.0 datasets contained full-length article SGML, including artwork files in Web-enabled graphic format. Version DTD 5.0 is now under development but unlike all other DTD versions, 5.0 will be XML-based and MathML-enabled.

Conclusions

This section describes the production processes that take place at Elsevier Science. A later section dealing with prototype development will describe what is necessary to move Elsevier data from the end point of the publisher's production systems (i.e., the CDs containing metadata and other content) into the prototype archive. It rapidly became clear that the bridge between the two worlds — that of the publisher and that of the archivist — is a very shaky one. While recognizing that much work has been done with such emerging standards as METS,[21] OAI, OAIS, etc., no archiving standards have yet been universally adopted jointly by major publishers and the academic community. The adoption of such standards is critical to the success of long-term electronic archives, but such standards urgently need further development and testing in a collaborative approach between the academic and publishing communities before they are likely to become an integral part of publishers' workflows.

As the study progressed, it became clear that the lack of accepted standards and protocols to govern such facets of metadata as transmission, data elements, format, etc., would be a major impediment in the future, not only to expanding the prototype but also to realizing the full potential of digital archives.

Major areas of concern regarding the present situation include generic problems associated with introducing "bridgeware" for any given publisher, as well as the unnecessary, if not prohibitive, costs and operational problems associated with developing, maintaining, and operating multiple different sets of bridgeware to accommodate different publishers having different metadata content and formats. This chaotic scenario has the potential to consume inordinate amounts of time and resources to produce, operate, and maintain such multiple instances of bridgeware.

Relating this chaotic environment to our experience in developing the prototype, the team observed that the lack of a commonly agreed upon set of metadata content necessitated the development of data replication and transformation software to convert the data received from the publisher (in Elsevier's case, in EFFECT format) into the format used in the archive itself (in our prototype, OAI and Dublin Core). At Yale, this transformation was made possible by using Extensible Style Language Transformation (XSLT) technology. However, our prototype development represents a far-from-optimal scenario. While it works in the case of Elsevier because the Dublin Core metadata can be generated from the information in the EFFECT datasets, there is no guarantee this would be true for every publisher.

A further danger is that this situation tends to lead to a fragmented approach in which archive information that is additional to what is needed for a publisher's current content operation will be added either as an afterthought at the publisher's end or as a pre-ingestion stand-alone process at the archive's end. However, based on Elsevier's past experiences in TULIP and in the early days of EES, the team observed that the necessary or additional metadata cannot be effectively and satisfactorily produced either as an afterthought post-production process on the publisher's side or as a pre-ingestion conversion activity at the archive's end. Approaching e-archiving in this fashion leads to distribution delays and a more complex production and distribution scenario, with all the accompanying potential to introduce production delays and errors.

The approach to adopt in creating electronic archives should be to recognize at the outset the needs of the archivist, i.e., to collect information to meet these needs as an integral part of the publishing and production process. Archival information will then be subject to the same level of production tracking and quality control as the information gathered in order to provide current-content service.

Creation of a Prototype Digital Archive

Introduction

In their seminal report of 1996, Waters and Garrett[22] lucidly defined the long-term challenges of preserving digital content. Ironically, the questions and issues that swirl about this new and perplexing activity can be relatively simply characterized as a problem of integrity of the artifact. Unlike paper artifacts such as printed scholarly journals which are inherently immutable, digital objects such as electronic journals are not only mutable but can also be modified or transformed without generating any evidence of change. It is the mutable nature of digital information objects that represents one of the principal obstacles to the creation of archives for their long-term storage and preservation. In the CPA/RLG report, Waters and Garrett identified five attributes that define integrity for a digital object: 1) content, 2) fixity, 3) reference, 4) provenance, and 5) context. At a different level, these attributes also represent various impediments or problems to the creation of digital archives. For example, the attribute called context involves, among other issues, the technological environment used to store, search, and render digital objects. Solutions to technological obsolescence such as refreshment,

emulation, or migration, are imperfect and can leave digital objects inaccessible — or even worse, in an irrevocably corrupted state — if they are not implemented correctly.

Unfortunately, there now exist numerous examples of data that are no longer available to scholars because there is no means to access information stored on obsolete computer hardware and software. In short, technological obsolescence is a vector that can, in a blink of the eye, undermine the integrity of all digital objects in an archive. The YEA was designed to learn how current models, standards, and formats could effectively address the principal problems of integrity and technological obsolescence that threaten the ontology of digital objects. What follows is a narrative that discusses the creation of the YEA, and the findings and lessons learned from our experiment.

Site visits to explore ongoing archival projects

Our first step in creating the YEA prototype was to learn about the publisher's technical systems and processes used to create electronic content. Knowledge of the publisher's workflows was important because this knowledge was needed to build an understanding, as defined by the attributes noted above, of digital objects produced by Elsevier Science's EW. At a minimum, the team needed to understand the structure and format of Elsevier Science content and the context or the technical environment needed to read and render the data. In addition, team members needed an understanding of the possible preservation metadata incorporated into Elsevier's data in order to address other issues such as the provenance and fixity of the digital objects that were to be stored in the YEA archive.

Next, team members conducted a review of research projects involved with electronic archives. The findings of this review showed that internationally the Open Archival Information System (OAIS) has been fast-tracked as a model that can be used to describe and create an infrastructure to support the activities of a digital archive. The OAIS reference model specifies and defines 1) an information object taxonomy and 2) a functional model for defining the processes of a digital archival system.[23] The team's other important finding was that the Open Archives Initiative (OAI),[24] an open source model for Web publishing, could be used as a means to access digital objects from the archive.

Archival projects that have adopted the OAIS model include the British Library in London, England; the National Library of the Netherlands, Koninkliijke Bibliotheek (KB) in Amsterdam, the Netherlands; and Harvard University in Cambridge, Massachusetts. In addition, WGBH, a non-profit media organization based in Boston, Massachusetts, plans to build its digital archive based on the OAIS model. The project's technical team decided to make site visits to the European projects and based upon the recommendation of Elsevier Science, to WGBH. Elsevier's technical staff also chose to visit the European sites, both because of Elsevier's relationship with these institutions and because the libraries are exploring the archiving of Elsevier Science content. In addition, both national libraries had already made significant contributions to the study of digital archives. The KB is known for its work on the NEDLIB project[25] and the British Library for its work on preservation description metadata. The site visits proved

important and valuable because they provided Elsevier with an opportunity to validate Yale's recommendation that the YEA prototype should be an implementation of the OAIS model. In addition, the team used these visits as a means to explore lessons learned from ongoing digital archive projects. The team wanted to take advantage of project successes and to avoid pitfalls that electronic archiving projects might have encountered. The elements that can contribute to the success or failure of electronic archiving are summarized as follows.

The archival projects at the British Library and the KB represent a continuum of results. The KB has been, by its definition, successful with its project while the British Library has experienced some significant problems with its plans to build a digital archive. Two factors account for the different outcomes. First, the British Library attempted to specify and develop production systems without the benefit of prototype models. In contrast, the KB is developing their archive from a series of research projects that have focused upon building prototypes of future production systems. The other success factor was related to the technical partnership between the library and the vendor chosen to develop the archival systems. In contrast to the British Library, which depended on the vendor to specify application requirements, the KB was able to specify application requirements to the vendor who in turn was able to develop system software. For example, the KB and IBM have successfully begun to modify IBM's Content Manager to integrate into the OAIS model. The KB and IBM have also successfully implemented a data management process and storage management system for digital objects. In addition, the KB and IBM have created beta code for an OAIS ingestion system. Because the KB possessed a deep intellectual understanding of the OAIS model and its requirements, staff were able successfully to exploit IBM's technical know-how. The other valuable lesson came from our site visit to Boston-based WGBH which has successfully partnered with Sun Microsystems to build a digital archive. Through this mutually beneficial partnership, Sun developed a new type of file system for large multimedia objects based upon content owned by WGBH. In return, WGBH was given hardware to support their digital archive. This experience reinforces the value of relevant partnerships that can, among other things, contain development costs.

These site visits also served to validate for YEA that the OAIS and OAI models could serve as foundations for a prototype archive. Consequently, the next objective for the study team was to build small but meaningful components of an OAIS archive. The team speculated, based upon knowledge of Elsevier Science content and Elsevier's EFFECT distribution standard, that rudimentary ingestion, data management, and archival storage processes could be created. Once built, these processes could be engaged to transform a primitive submission information package (SIP) into a primitive archival information package (AIP), which in turn could be searched and rendered via an OAI interface. The prototype-building work began in August 2001 and was successfully completed four months later in December 2001.

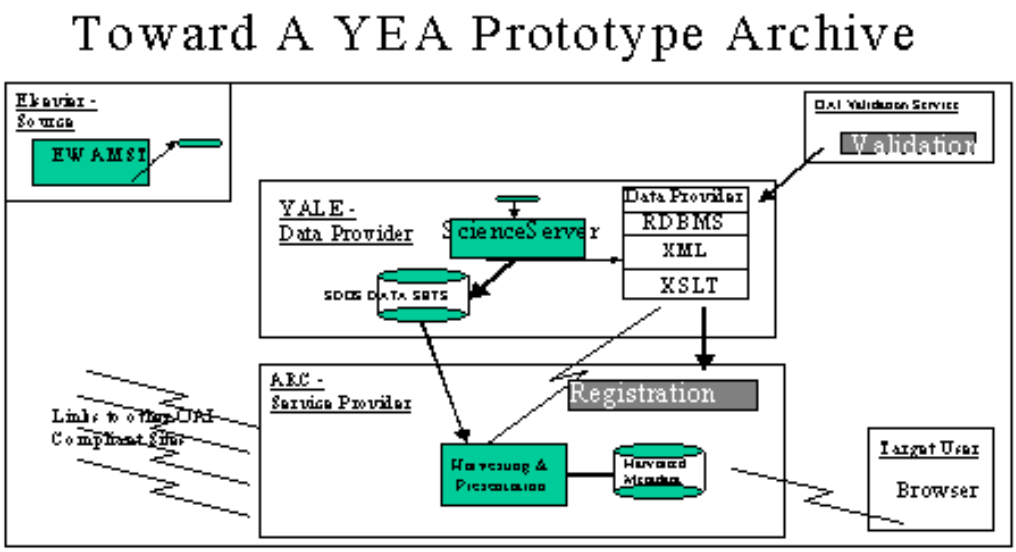
In May 2001, sandwiched between the site visit to WGBH and the ground-breaking for the YEA prototype, the team hosted a presentation on digital archiving from the J. P. Morgan Chase I-Solutions group. Chase's I-Solutions potentially offers a means to reduce

the total cost of ownership of an archive for a library. That is, Chase's I-Solutions provides archival storage services to commercial businesses that are required to store business transaction data, such as checks and loan agreements, for a minimum period of seven years. Potentially, scholarly archives could outsource the storage management function of their archives to large commercial institutions such as Chase. The Yale team speculated that perhaps, by taking advantage of the scale offered by the Chase archival infrastructure, it might be possible to reduce the total unit cost of storage for a digital object. However, it is difficult to tell at this time whether the cost-saving could be passed on to the customer. While the economics of this concept deserve additional research, other issues make Chase's I-Solutions less useful at this point in time. These issues revolve around trust and major differences between the nature of commercial and scholarly archives. Would the academic community trust a very large commercial institution to archive their content when this community has reservations about relying upon publishers to archive their content? In addition, the I-Solutions archive is currently not designed to capture preservation metadata with stored digital objects. Equally important, the commercial archive is designed to store content for a short period, not the one hundred plus years that academic archives require. Access to the stored objects also poses a problem. Once archived with Chase, an object or its metadata cannot not be altered. Accordingly, at this stage the team believed it was premature to establish a collaborative relationship with Chase.

The Concept of the YEA prototype Archive

The Yale/Elsevier technical team perceived that the EFFECT standard and the OAI protocol could be fashioned to build two components of the OAIS model. That is, data in Elsevier's EFFECT format could be transformed into a Submission Information Package (SIP) and the data provider component of the OAI model could be used as an OAIS data management process. Archival storage for the digital objects could be created with a standard UNIX (Solaris) file system. However, the design of the prototype hinged upon the team's ability to convert the EFFECT metadata found in the publisher's SDOS distribution datasets into an XML format. Once converted to XML, the metadata could be transformed into the Dublin Core format that is a requirement for the OAI data provider protocol. Thereafter, the converted metadata could be harvested and exposed by an OAI compliant service provider. Figure 1 below provides a schematic of the YEA prototype.

Figure 1



The Hardware and Software Infrastructure

The YEA prototype is split across two different hardware and software platforms. This is idiosyncratic and was done for the sake of expediency. The OAI components are deployed on an IBM Thinkpad model T20 that runs Windows Professional 2000. The Thinkpad has one Intel x86 based, 700 Mhz processor and about 400 Mbytes of memory. The PC's internal disk storage drive has a capacity of 11 Gbytes. The OAIS archival storage component is hosted on a Sun Enterprise 450 workgroup server that run Solaris version 2.8. The system has two (Sun Ultra Sparc-II) 450 Mhz processors and 2 Gbytes of memory. An external Sun A1000 disk array, which provides about 436 Gbytes of storage, is attached to the server.

The OAI data provider software was developed at the University of Illinois at Urbana-Champaign as part of another Mellon foundation project. To run the application the following software is required:

- Microsoft 2000 Professional
- Microsoft Internet Information Server version 4.0 or higher
- Microsoft XML Parser (MSXML) version 4.0
- Microsoft Access

The OAI data store from the University of Illinois-UC conforms to version 1.1 of the OAI protocol. The data service has a simple but powerful design that separates XML metadata files from OAI protocol administrative information about each object.

Individual XML files are stored in a Windows file system, and the OAI administrative data are stored in four simple RDBMS (MS Access) tables. In the YEA implementation, the XML files are stored in a MS Windows file system, and the administrative data is stored in an MS Access database. Data in these tables contain the substance of the protocol, without which the data server could not respond to an OAI command or request for information. For example, the metadata table contains all the XML information necessary to describe a metadata object. This information includes the XML namespace, schema, and style sheet information that is needed to transform tagged data in the XML files to Dublin Core. The purpose of the Object table in the database is to maintain a set of pointers to these files. The Repository table provides descriptive information about the repository and the Set table defines collects that are within the repository.

The OAI service provider software is not part of the local environment but was developed at Old Dominion University. The search service called "Arc" is an experimental research service of the Digital Library Research group at Old Dominion University. Arc is used to harvest OAI-compliant repositories, making them accessible through a unified search interface over the network.

Programmatic Process

Elsevier Science distributes electronic content in a proprietary format that has the fundamental components or structures of a SIP. The standard includes a data object and representation information to provide meaning to the bits that compose the object. As noted above, Elsevier's SDOS datasets are distributed in EFFECT format and contain data objects that are composed of text, image, and PDF files. The encoding standards for each object type are identified and partially defined in the EFFECT dataset. In addition to containing these objects, the EFFECT standard also contains some preservation description information metadata that are needed to maintain the long-term integrity of an archived information object. Finally, the EFFECT standard contains packaging information that 1) uniquely identifies the dataset that contains the information objects so that they can be located through a search process, and 2) provides identification and descriptive information about the delivery vehicle, i.e., the CD-ROM or FTP protocol used to transport the content to a customer or archive. Other representation information in the standard defines the directory structure of the distribution dataset and a logical mapping of how data objects can be assembled to display the digital object through rendering software.

As delivered to Yale, the usefulness of the SDOS data for archival purposes was diminished because these data are not in a normative format such as XML. To transform these data into an XML format, the researchers adopted a piece of software from Endeavor Information Systems (EIS), another company owned by Elsevier Science. EIS was able to process an EFFECT dataset and convert the metadata into XML. Thereafter, the metadata had to be converted to the Dublin Core standard so it would comply with the OAI protocol. Using the expertise of a librarian, the team mapped the EFFECT metadata tags to qualified Dublin Core tags. Once these tags were mapped, an XSLT style sheet was created to dynamically transform EFFECT tags to Dublin Core in response to an

OAI request. Approximately fifty items from an SDOS dataset were converted and loaded into the OAI database.

Once data were loaded, the implementation was validated with a tool provided by the Digital Library Research group at Old Dominion University. The tool is called the Open Archives Initiative Repository Explorer.[26] Once the contents are validated, the repository is certified to understand any verb or command in the OAI protocol. The YEA prototype passed this implementation test. After making small modifications to the main application module for the data provider service, the team could experiment with initiating OAI commands to the YEA database. However, the search interface to these data was limited and lacked a feature for retrieving and displaying the digital objects that reside in archival storage.

Figure 2

To provide a unified search interface to YEA, Elsevier Science gave the planning team permission to register the archive with the Arc service provider. Shortly thereafter, Arc's Web harvesting daemon extracted metadata from YEA and, in addition to other repositories, the YEA officially became a repository that could be searched through the Arc interface. Now, not only could YEA metadata be exposed in Dublin Core format, but also links to digital objects in archival storage became active. A reader would not only see metadata about Elsevier Science but could also extract and view the actual content in PDF format through a simple Web browser.

View 1


arc
Cross Archive Searching Service
Simple search Advanced Search Register Browse Help
Matches were found in these archives
archive Hits
yea 1
Search Summary
Search Summary
Search: where contains(creator, 'Schreiber',1)>0 and archive = 'yea' Sort: order by score(1) desc
Group By: archive Value:yea
SEARCH RESULTS
Title
A comprehensive electromotive force series of redox couples in soda–lime–silicate glass
Creators
Henry D. Schreiber Department of Chemistry, Virginia Military Institute, Lexington, VA 24450-0304, USA | Nicholas R. Wilk Jr. Department of Chemistry, Virginia Military Institute, Lexington, VA 24450-0304, USA | Charlotte W. Schreiber Department of Chemistry, Virginia Military Institute, Lexington, VA 24450-0304, USA
Description
The redox equilibria of several multivalent elements were measured in a soda–lime–silicate composition at 1400°C to determine the relative reduction potentials for Co^{2+}–Co^{0}, Cr^{6+}–Cr^{3+}, Cr^{3+}–Cr^{2+}, E
Archive
yea
Date Stamp
2001-10-24
Document ID
oai:yea:185559

View 2

Source DL:	yea
Identifier	http://epreservation.library.yale.edu/YEA/OHL0508A/00223093/V0253I01/00004428.pdf
Title:	A comprehensive electromotive force series of redox couples in soda–lime–silicate glass
Creator:	Henry D. Schreiber Department of Chemistry, Virginia Military Institute, Lexington, VA 24450-0304, USA
Creator:	Nicholas R. Wilk Jr. Department of Chemistry, Virginia Military Institute, Lexington, VA 24450-0304, USA
Creator:	Charlotte W. Schreiber Department of Chemistry, Virginia Military Institute, Lexington, VA 24450-0304, USA
Subject:	E160 G210 O160 S170
Description:	The redox equilibria of several multivalent elements were measured in a soda–lime–silicate composition at 1400°C to determine the relative reduction potentials for Co^{2+}–Co^{0}, Cr^{6+}–Cr^{3+}, Cr^{3+}–Cr^{2+}, Eu^{3+}–Eu^{2+}, Fe^{3+}–Fe^{2+}, Mn^{3+}–Mn^{2+}, Ni^{2+}–Ni^{0}, and Ti^{4+}–Ti^{3+} in this solvent. A comprehensive electromotive force (emf) series of redox couples for soda–lime–silicate glass at 1400°C was then developed by calibrating these selected reduction potentials to the corresponding emf series in a reference alkali borosilicate composition. The reliability of this reference emf series was corroborated by measurements of the reduction potentials of redox couples by electrochemical methods. Applications of the comprehensive emf series include (1) the estimation of the redox ratio of a specific multivalent element under a prescribed set of processing conditions, (2) the identification of whether a multivalent element acts as an oxidizing or reducing agent with respect to another redox couple in the glass-forming system, (3) the prediction of color and other redox-dependent properties of the glass product, and (4) an explanation for the fining characteristics of certain polyvalent elements.
Format:	Article Full-text (PDF)
Language:	EN
OAI Info:	
ID:	oai:yea:185559
Set:	YEA E-Journal Archive
DateStamp:	2001-10-24

View 3

ELSEVIER

Journal of Non-Crystalline Solids 253 (1999) 68–75

JOURNAL OF NON-CRYSTALLINE SOLIDS

www.elsevier.com/locate/jnoncrysol

A comprehensive electromotive force series of redox couples in soda–lime–silicate glass

Henry D. Schreiber *, Nicholas R. Wilk Jr., Charlotte W. Schreiber

Department of Chemistry, Virginia Military Institute, Lexington, VA 24450-0304, USA

Abstract

The redox equilibria of several multivalent elements were measured in a soda–lime–silicate composition at 1400°C to determine the relative reduction potentials for Co^{2+}–Co^{0}, Cr^{6+}–Cr^{3+}, Cr^{3+}–Cr^{2+}, Eu^{3+}–Eu^{2+}, Fe^{3+}–Fe^{2+}, Mn^{3+}–Mn^{2+}, Ni^{2+}–Ni^{0}, and Ti^{4+}–Ti^{3+} in this solvent. A comprehensive electromotive force (emf) series of redox couples for soda–lime–silicate glass at 1400°C was then developed by calibrating these selected reduction potentials to the corresponding emf series in a reference alkali borosilicate composition. The reliability of this reference emf series was corroborated by measurements of the reduction potentials of redox couples by electrochemical methods. Applications of the comprehensive emf series include (1) the estimation of the redox ratio of a specific multivalent element under a prescribed set of processing conditions, (2) the identification of whether a multivalent element acts as an oxidizing or reducing agent with respect to another redox couple in the glass-forming system, (3) the prediction of color and other redox-dependent properties of the glass product, and (4) an explanation for the fining characteristics of certain polyvalent elements.

1. Introduction

affect the stabilization of such redox states in the

Lessons Learned and Next Steps

The ability to preserve digital information for long periods of time is dependent upon the effectiveness of the models, standards, formats, and protocols that can be applied to overcome problems associated with technological obsolescence and the maintenance of the integrity of digital objects. The YEA prototype provided some evidence that the OAIS and OAI models can be used to create such an archive. The OAIS model specifies metadata that is needed to preserve the integrity of digital information for long periods of time.

The e-archiving planning team also learned that some of this metadata, e.g., the Preservation Description Information, is already incorporated into distribution datasets disseminated by at least one major publisher, Elsevier Science. However, SIP elements included in Elsevier's EFFECT standard need formally to be developed to include all metadata concerning the fixity, reference, provenance, and context of a digital object. The team recognizes that to obtain these data, production processes used to create Elsevier's content will have to be modified. To be helpful to the archival community, Elsevier needs to assure that its metadata conforms to emerging standards for SIPs.

Recently, Harvard University released a specification for a SIP expressed in XML format.[27] The effectiveness and usefulness of this model can only be determined through robust testing from real applications. Elsevier Science has plans to change its workflows so that content and distribution standards are XML-based. Such a change could provide an opportunity for the planning team, in Phase Two, to work to develop an advanced prototype SIP that conforms to the Harvard specification.

The team's experience with OAI data provider software showed promise for rapid deployment. The economic advantage of using open source software to sustain archives is self-evident. The OAI protocol also showed promise for interoperating with the OAIS model. One shortcoming of a standard OAI implementation is that the Dublin Core standard does not allow all of the rich metadata found in an Information Package to become manifest. Dublin Core was developed as a low-barrier means of searching for data; consequently, it has a limited tag vocabulary. However, the OAI standard does allow for multiple manifestations of an object's metadata. In Phase Two, the Yale-Elsevier team will consider how best to develop an XML schema that can permit the exposure of all metadata found in a formal AIP.

The simplicity of the OAI data store that was implemented suggests that the data management architecture is also scalable and portable. The project team has experiential evidence suggesting that applications that store data objects external to a RDBMS are more flexible, easier to maintain, and more portable across operating systems. These features can potentially make migrations less problematic and more economical when they need to occur.

As alluded to elsewhere in this report, advances in the creation of digital archives are dependent upon systems that can interoperate, and on data and content that are portable. The suite of XML technologies that includes XML, XSL, XPATH, and XSLT provides some hope and high expectations that electronic archives can be successfully architected and implemented. The fact that the project team was able quickly to create an XSLT script with the use of a XML tool demonstrates the promise of already-available technologies. In addition, there is substantial evidence that text data in tagged format such as SGML or XML is very portable across different platforms. The challenge to e-archives is how to make other data formats such as audio and video equally portable across systems. This challenge will be particularly important for specification of standards for Dissemination Information Packages (DIPs).

Finally, the continued success of this impressive beginning between the Yale Library and Elsevier Science to explore digital archives hinges upon a few critical success factors. First is Elsevier Science's continued and unflagging support of Yale Library as a strategic planning partner. The professional relationships and confidence that each institution now has in the other is invaluable. Together, the two institutions are poised to become leaders in the electronic archiving field. Second, Yale and Elsevier need to gain experience with other publishers' formats. In this, Elsevier Science can be helpful, because many of its rendering applications load data from other publishers. Finally, Elsevier Science is interested in exploring the concept of an archival standard that can apply across

publishers. Their support and resources could contribute significantly to the development of these systems and standards.

Digital Library Infrastructure at Yale

Existing Infrastructure

At present, the Yale University Library has several successful components of a Digital Library Infrastructure (DLI) in production use and is actively pursuing additional key initiatives. The Luna Insight image delivery system, for example, elegantly supports faculty and student use of digital images for teaching and study, but it does not yet address the long-term preservation of those images. Over the next three years, we have committed to a substantial expansion of the Luna imaging initiative supported by grant funding from the Getty Foundation, digital collection development in the Beinecke Library, and increased dedication of internal Library resources to the project. Our Finding Aids database provides access to archival finding aids encoded in formats that will endure over time, but, like many of our current stand-alone systems, this resource is not yet tightly integrated into an overall architecture. The participants in the Finding Aids project (including the Beinecke Library, the Divinity Library, the Music Library, and Manuscripts and Archives) continue to add new material to this database and to enhance the public interface.

The Orbis online catalog is perhaps the best example of a system designed, operated, maintained, and migrated with a focus on long-term permanence. The Yale Library staff are currently engaged in the process of migrating from the NOTIS library management system to the Endeavor Voyager system, with production implementation scheduled for July 2002. We are confident that the bulk of our retrospective card catalog conversion effort will also be finished by that date. The meticulous examination of hardware options for the LMS installation has resulted in a high level of expertise among Yale systems staff in the technical requirements for robust, large-scale processing and storage of critical data. The completion of these two fundamental and resource-intensive projects will position the Library well for a substantial investment of energy in innovative digital library initiatives.

In order to enhance reader navigation of digital resources, for instance, we have purchased SFX and MetaLib from Ex Libris. SFX context-sensitive reference linking went into production use at Yale in January 2002 for four information providers: Web of Science, EbscoHost, OCLC FirstSearch, and the OVID databases. Immediately following the Voyager implementation next summer, library staff expect to implement the MetaLib universal gateway in order to deliver unified access, federated searching, and personalized services across a wide range of local and remote resources.

Integration of Library applications with the wider University technical environment is another key principle guiding our work. The Library, for instance, is well advanced in utilizing the campus-wide authentication infrastructure based upon a Central Authentication Service and a universal "NETID." Other areas for collaborative activity

include the University Portal project (now in preliminary planning stages), the Alliance for Lifelong Learning,[28] and the Center for Media Initiatives (faculty technology support).[29]

Current Commitments

University Electronic Records Management Service

The Library's University Archives program provides records management services for University records of enduring value. Yale e-records comprise those digital objects (such as e-mail, administrative databases, Web sites, and workstation products) that are created primarily to administer and sustain the business of the University. The University is committed to preservation and access for its record in perpetuity through the services of the Library's Manuscripts and Archives Department. As part of its tercentennial initiatives, Yale augmented significantly the Library's records management capabilities. Recently received permanent funding for a comprehensive University Archives Program will help to ensure that preservation of University records in digital formats is properly addressed. New attention to the management of the University's electronic records is a necessary and natural extension of the responsibilities now fulfilled for paper records. The service (ERMS) will develop the capacity to ensure the long-term preservation of and access to vital University records, while ensuring their intellectual integrity and guaranteeing their authenticity. The service will harvest and augment the metadata of the records to enhance access and support data management, migration, and long-term preservation. The ERMS will also provide consultation and guidance to University departments implementing electronic document management systems, to ensure that such systems meet the needs of the University for long-term preservation of and access to their content.

Beinecke Rare Book and Manuscript Library Digital Collections Project

The Beinecke Library is now (January 2002) embarking on a major multi-year digitization project which will generate approximately two thousand images per month from book and manuscript material in the Beinecke collections. Luna Imaging, Inc., is acting as consultant for the newly-created scanning facility, and Beinecke staff are committed to high standards for image quality, metadata creation, and long-term archiving. Discussion has already begun among Beinecke staff, the Library Systems Office, and ITS staff about the appropriate mechanisms for ingestion of this material into the Yale Electronic Archive.

Next Steps for Establishing the YEA

In December 2001, the Collection Management Goal Group (CMGG), one of six strategic planning groups created by the new University Librarian, strongly recommended that the University "establish a Digital Library Infrastructure (DLI) in the Yale University Library worthy of an institution with a 300-year history of acquiring, preserving, and providing access to scholarly research material." There was a powerful

conviction in this goal group, and among those consulted by the goal group, that digital preservation is the highest-priority unmet need in the Library's nascent digital infrastructure. The CMGG summarized the first-year objectives as follows:

1. Establish the scope of need for a local digital archive (the metaphor used is that of a "Digital Library Shelving Facility," based on the Library's high-density shelving service near the campus).

2. Establish an appropriate infrastructure for a digital archive with a focus on the Mellon-funded project to preserve scholarly electronic journals.

3. Important potential candidates for preservation include:

- Unique digital material acquired by Yale (e.g., literary archives in the Beinecke Library)

- Born-digital material acquired/selected by Yale, not adequately preserved elsewhere (e.g., selected government documents, works published online, Web sites)

- Formal partnerships with providers (e.g., Elsevier e-journals)

- Surrogates (preservation) created/acquired (e.g., digital reformatting of brittle books)

- Surrogates (public access collections) created/acquired (e.g., teaching images collected or created by the Visual Resources Collection and the Beinecke Rare Books Library)

4. Establish metadata requirements for materials destined to reside in the Archive.

Senior Library managers are committed to the aggressive pursuit of funding for the creation of this essential digital infrastructure.

Endnotes

[1] Stewart Brand, *How Buildings Learn: What Happens After They're Built* (NY: Viking, 1994).

[2] Under the auspices of the International Union of Pure and Applied Physics (IUPAP) (http://iupap.org), a workshop on the Long-Term Archiving of Digital Documents in Physics was held in Lyon, France, 5-6 November 2001. As the digitization of physics literature has become increasingly widespread, concern about the long-term preservation of this literature has risen. The workshop brought together society and commercial publishers, librarians, and scientists to discuss issues related to the long-term archiving of electronic publications. See http://publish.aps.org/IUPAP/.

[3] Carol Fleishauer is Associate Director for Collection Services of the MIT Libraries and a founding member of the NERL consortium.

[4] 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.

[5] *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*, Publication 77 (Washington, DC: Council on Library and Information Resources, January 1999), online at http://www.clir.org/pubs/reports/rothenberg/contents.html.

[6] 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). Online at ftp://ftp.rlg.org/pub/archtf/final-report.pdf.

[7] For additional information about OCLC's programs in the digital preservation arena, see: http://www.oclc.org/digitalpreservation/.

[8] For a brief overview of the NDIIPP, see Deanna Marcum, "A National Plan for Digital Preservation: What Does it Mean for the Library Community," CLIR Issues 25 (January/February 2002): 1, 4. Online at http://www.clir.org/pubs/issues/issues25.html#plan.

[9] John C. Bennett, "A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Material," British Library Research and Innovation Report 50, Version 1.1, JISC/NPO Studies on the Preservation of Electronic Materials (London: British Library Research and Innovation Centre, 23 June 1999). Online at http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/rept011.pdf.

[10] Anne J. Gilliland-Swetland and Philip B. Eppard, "Preserving the Authenticity of Contingent Digital Objects: the InterPARES Project." *D-Lib Magazine* 6.7/8 (July/August 2000). Online at http://www.dlib.org/dlib/july00/eppard/07eppard.html.

[11] Marie-France Plassard, ed., IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report*, UBCIM Publications - New Series Vol 19 (München: K. G. Saur, 1998). Online at http://www.ifla.org/VII/s13/frbr/frbr.htm.

[12] Cited above at note 4.

[13] "The Making of America II Testbed Project White Paper," Version 2.0 (15 September 1998). Online at: http://sunsite.berkeley.edu/moa2/wp-v2.pdf.

[14] OCLC/RLG Working Group on Preservation Metadata, "Preservation Metadata for Digital Objects: a Review of the State of the Art" (15 January 2001). Online at http://www.oclc.org/research/projects/pmwg/presmeta_wp.pdf.

[15] The complete archive of Elsevier Science SGML/XML DTDs is available online at http://support.sciencedirect.com/tectext_sgml.shtml.

[16] This list can be found at http://dublincore.org/documents/2000/07/11/dcmi-type-vocabulary/.

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

[18] For information on OAI, see http://www.openarchives.org/.

[19] See Marthyn Borghuis, *et al*, *Tulip: Final Report* (New York: Elsevier Science, 1996). Online at http://www.elsevier.nl/homepage/about/resproj/trmenu.htm.

[20] To learn more about EFFECT, see http://support.sciencedirect.com/tecindex.htm.

[21] The Metadata Encoding & Transmission Standard (http://www.loc.gov/standards/mets).

[22] Previously cited at note 7 above.

[23] Previously cited at note 4 above.

[24] Previously cited at note 19 above.

[25] NEDLIB is a collaborative project of European national libraries. It aims to construct the basic infrastructure upon which a networked European deposit library can be built. The objectives of NEDLIB concur with the mission of national deposit libraries

to ensure that electronic publications of the present can be used now and in the future. See extensive documentation at http://www.kb.nl/coop/nedlib/homeflash.html.

[26] This site presents an interface to interactively test archives for compliance with the OAI Protocol for Metadata Harvesting. See http://oai.dlib.vt.edu/cgi-bin/Explorer/oai1.1/testoai.

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

[28] The Alliance for Lifelong Learning is a joint venture between Oxford, Stanford, and Yale Universities. See http://www.allianceforlifelonglearning.org/.

[29] For additional information, see http://cmi.yale.edu/.

Part III: Appendices

Appendix: Three Models of Archival Agents

	Scope of archival commitment	Users	Possession of content	Technical environment	Access	Interlibrary loan	Distribution of costs
1 – De facto archival agent (e.g., OhioLink, CSIRO)	Made to the readers using the archive or to consortium members	Users are primarily bench scientists with information needs similar to those of other users of the publisher's online content	Perpetual license is the functional equivalent of subscriber ownership of content, but possession is governed by contact law (i.e., the license) rather than copyright law	Local load of content; rendering software identical to that ordinarily provided by the publisher	License imposes few restrictions for authorized users; no online access allowed to others unless a separate archival agreement is reached (at which time the de facto archival agent may become a self-designated archival agent)	As a ScienceDirectOnSite licensee, may provide interlibrary loan services by creating paper output from online content	Costs borne by licensing library or distributed among consortium members
2 – Self-designated archival agent (e.g., National Library of the Netherlands)	Made to readers or to libraries, generally within a given geo-political unit	Same as above, except that national libraries would likely have relatively few bench scientists working onsite	Same as above	Same as above	License imposes few restrictions as regards readers physically present at the archival agent's service sites; no online access allowed to libraries beyond the archival agent except, temporarily, in the case of service disasters the publisher is otherwise unable to handle	Same as above	Borne solely by the archival agent, which may be a government agency; copyright deposit laws may be relevant
3 – Publisher-archival agent partnership (e.g., Yale/Elsevier	Made to the publishers and, through the publisher, to readers and	Users are primarily not bench scientists but those pursuing	Same as above	Local load of content; metadata, storage technology,	License sets terms similar to those for self-designated archival agents	Interlibrary loan for commercially viable content is not part of the agent's mission; few restrictions on	Borne by the archival agent; publisher may wish to

partnership)	libraries that do not benefit from the commitments made by other archival agents	the historical inquiries typically supported by archival services; ordinary users of the publisher's content that is no longer commercially viable are also supported by the archive		and rendering software may differ from that ordinarily provided by the publisher, given the different audience of users	for the commercially viable part of the publisher's content; few restrictions on access to the content that it not commercially viable; protection against service disasters is a possible, but secondary, mission	access to the content that is no longer commercially viable	subsidize archival activity

NOTES

NOTE 1: A library willing to negotiate a national site license could function for an entire nation in the manner of a consortial archival agent.

NOTE 2: The access column defines the scope of library operational security provided by each archival agency. For instance, the consortial archival agent ensures permanent online access to all of the publisher's content but only to consortial members, while the publisher-archival agent partnership ensures largely unrestricted permanent online access only to commercially non-viable content.

NOTE 3: The defining concern of the publisher-archival agent partnership, as distinguished from the two other kinds of archival agent, is to identify content that is not commercially viable. Such an agent seeks in this way to minimize conflicts between the commercial mission of the publisher and the preservation/access mission of the archival agent. No boundary between commercially viable and non-viable content has yet been identified. Such boundaries may in time be established by a "rule of thumb" (as has been done with JSTOR or proposed in the case of PubMed Central), or by invoking some set of yet-to-be-specified "trigger events."

NOTE 4: Self-identified archival agents and publisher-archival agent partnerships might provide access to commercially viable content to libraries that once had licenses for that content and/or to libraries in developing countries to which publishers wish, as a matter of good public policy, to provide content for free or at steeply discounted prices.

NOTE 5: Should boundaries between commercially viable and non-viable content be established, consortial and self-identified archival agents may wish to reshape their services with regard to those boundaries, to provide some online access to readers beyond the boundaries of the consortium or physical site. If enough archival agents did this, it is not clear that the publisher-archival agent partnership would have a distinctive function.

NOTE 6: One reason for the maintenance of paper publications is our present inability to identify a boundary between commercially viable and non-viable content. The costs associated with these subscriptions therefore represent, in part, the preservation costs libraries bear. Put differently, the cost burdens of paper publication result, in part, from libraries' unwillingness to put their preservation mission at risk and from publishers' unwillingness to put their commercial mission at risk. This impasse exists because, for online information resources, libraries do not own any information carrier that can be preserved.

Appendix

Description of Metadata Elements for the Yale Electronic Archive

We gratefully acknowledge our indebtedness to the British Library for sharing an April 2001 draft version of their *Description of Metadata Elements for the Digital Library System* document on which this work is largely based.

29 November 2001

Table of Contents

    - PrimaryTitle
    - AlternativeTitle
    - SeriesTitle
    - SeriesTitleNumber
    - ArticleTitle
    - UniformTitle
- Subject Group
    - LCSH
    - NameAsSubject
    - FreeText
    - Other
- Description Group
    - Abstract
    - FreeText
    - PubStatus

Coverage Group

- Date Group
    - DateIssued
    - DateReceived
    - DateArchived
    - LicenceCheckDate
    - DateModified
    - DateCreated
    - VitalDate
    - LicenseStartDate
    - EventDate

Terms Group

- TermsOfAvailability Statement
- TermsOfAvailability Reference

Type and Identifier Group

-
    - ResourceType
    - ObjectType
    - ObjectPreservation Category
- ResourceIdentifier
    - System IDs
    - Descriptive IDs

Format Group
Relation Group

- RelationIsVersionOf
- RelationIsPartOf
- RelationTransformedfrom
- RelationReplacedby
- RelationReplaces
- RelationExternalObject
- RelationRefersTo
- RelationOtherArchive

History Group

- Custody History
- Ingest History
- Preservation History
    - ProcessName
    - ProcessDescription
    - ProcessSpecification
    - CriticalSoftware
    - ProcessResult
    - ProcessAgent
    - ProcessDate

Object Group

- DigitalSignature
- DigitalSignatureName
- OperatingEnvironment
- ObjectIdentifier

External Object Group

- RelationExternal Object
- RelatedInformation Object

Rights Information Group

- Rights Group
    - RightsStatement
    - RightsHolder
- Licence Group
    - LicenceType

SystemParameter Group

- OriginalCarrier

Appendices

1. Controlled vocabularies maintained by the archive

1.1 Agent Roles List
1.2 Resource Type List (note: we should use Dublin Core Metadata Initiative Resource List)
1.3 Object type List
1.4 Preservation Category list
1.5 Process Name list
1.6 Original Carrier List
1.7 Other Subject Vocabularies

Status Column key:
R = Repeatable
M = Mandatory (in order for object to load into ARCHIVE)
Mif = Mandatory under certain conditions

See http://www.diglib.org/preserve/yalemetadata.pdf for field definitions.

Appendix

1. Controlled vocabularies maintained by the archive

The following lists contain the initial listing of values that will be entered in elements where the value is selected from a drop down list. These list are extensible and further values will be added as the need is identified.

1.1 Agent Roles List

Agent Role List	***metadata category***
Author [aut]	Descriptive
Conference [cnf]	
Copyright holder [cph]	Descriptive
Correspondent [crp]	Descriptive
Editor [edt]	Descriptive
Illustrator [ill]	Descriptive
Licensee [lse]	Descriptive
Licensor [lso]	Descriptive
Publisher [pbl]	Descriptive
Reviewer [rev]	Descriptive
Speaker [spk]	Descriptive
Translator [trl]	Descriptive
Other [oth]	Descriptive
Archive Specific Roles	
Digitiser	Administrative
Custodian	Administrative
Preservation User	Administrative
RightsHolder	Administrative
Repository Name	Administrative

1.2 Resource Type List

(note: we should use Dublin Core Metadata Initiative Resource List)
Image,
Audio,
Video,
Multimedia,
Text,

Executable,
PDF,
SGML,
XML,
Dataset

1.3 Object type List

Map (+OS),
Sheet Music,
Media (inc. sound and video),
Pictorial,
Software,
Serial (inc. Newspapers),
Issue,
Article (FLA),
Letter (COR, DIS, SCO),
Review (book review BRV; product review PRV),
Advertisement (ADV),
Notices (publisher's note PUB),
Erratum (ERR),
Abstract (when published as separate item; ABS),
Addendum (ADD),
Announcement (ANN),
Calendar (Meetings Calendar CAL),
Editorial (EDI),
Alert (LIT),
News (NWS),
Contents (OCN),
Report (patent report PNT; personal report PRP),
Request (REQ),
Survey (SSU),
Miscellaneous (MIS).

1.4 Preservation Category list

Voluntary
Purchased
Contractual Arrangement

1.5 Process Name list

Scan of transparency
etc.

1.6 Original Carrier List

CD-ROM
DVD
DLT IV cartridge

Other
etc.

1.7 Other Subject Vocabularies

To be defined as needed
|

|Note: The information recorded for EFFECT and DTD equivalence is incomplete and provided only as a reference of the type of cross-mapping that can and should occur for proper ingest of publisher metadata.

Appendix
Elsevier Science Technical Systems and Processes

Glossary for Standards

Distributed content from the ES warehouse in the Netherlands contains data that have been encapsulated or bundled in five different distribution formats that reflect the technological advancement of ES production and distribution process. The distribution datasets were once called Elsevier Electronic Subscriptions (EES), now obsolete, and were replaced in 1998 by Science Direct OnSite (SDOS): The version history is as follows:

PRECAP: Pre-computer aided production; placed into service in 1995

CAP: Computer Aided production; placed into service in 1997

EES V1.0

TIFF files containing scanned images

Raw ASCII text files, one for each page

SGML citation files

Dataset.toc file in EFFECT 4.0 specification

EES Version 1.1

Same as above except that the TIFF image page files were replaced by wrapped PDF files that contained an editorial item

Dataset.toc file in EFFECT 4.0 specification

EES Version 1.2

Same as EE version 1.1 but editorial items could be contained in wrapped PDF or true PDF format, i.e., converted from original Postscript file -- highest resolution.

Dataset.toc file in EFFECT 4.0 specification

SDOS Version 2.0

PDF files containing an editorial item in wrapped or true format

Raw ASCII files containing an editorial item in wrapped or true format

SGML citation files containing bibliographic data for editorial items

Dataset.toc file in EFFECT 4.0 specification

SDOS Version 2.1

PDF files containing an editorial item in wrapped or true format

Raw ASCII files containing an editorial item in wrapped or true PDF format

SGML citation files containing bibliographic data for editorial items and article references in structured format

Dataset.toc file in EFFECT 4.0 specification

SDOS Version 3.0

PDF files containing a publication item in wrapped or true format

Raw ASCII files containing a publication item in wrapped or true format

Full article SGML files for publication items, artwork files in Web-enabled graphical formats

Dataset.toc file in EFFECT 4.1 specification

Data Components Found in EES and SDOS Datasets

Page Images

Black and white

TIFF 5.0 standard

Scanned at 300 dpi

Maximum scan is European A4, i.e., 210x297mm2

Compression ITU T.6, aka CCITT Fax group 4, for an average page 8%

Compression is achieved, i.e., 1M=+- 80Kbytes

White background and black characters

Raw Text Files

Each page image has a corresponding raw ASCII file

Produced from OCR procedures

No keyboarding/editing/spell-checking is performed on them

Contain only ASCII characters 32-126

Provided as a basis for searchable indexes -- not for end users

SGML Files

Text of editorial items

SGML files are encoded in plain ASCII

SGML files have two extension attributes: ".sgc" and ".sgm"

Former means SGML data for heading information and the latter means full SGML content

Note: SDOS2.1 contains only ".sgc" files

Other Files

Pertains to distribution of content. Supplier and receiver agree that files with these other formats for content can be packaged in SDOS 2.1 datasets.

Adobe Acrobat Portable Document Format (PDF) Item/Page basis. Item based files contain a one-to-one ratio of one PDF file for one issue article.

Page-based PDF files contain pages that are not part of a clearly identified item/article such as front and back covers, advertisements etc.

Together item-based and page based PDF files can be used to reconstruct the entire paper journal in electronic format.

True/Distilled: original typesetter Postscript files

- no paper scanning steps

- same quality as final paper journal issue

Wrapped: image scanning on the paper journal issue

- TIFF images - fax group 4 encapsulated in PDF code

- lesser quality then distilled

Encapsulated PostScript (EPS)

Joint Photographer Expert Group (JPEG) encoded files

Hypertext Markup Language files

Compuserve Graphics Interchange Format (GIF) compressed files

TEX encoded files

CHECKMD5.FIL: Checksum facility to ensure the validity or integrity of the data distributed to the Client.

EFFECT- DATASET.TOC FILE:

Contains all cross-indexing reference data needed to load into an application or database. See EFFECT document for general rules of this file.

DATASET.TOC is split up into records that are broken into four major divisions.

_t0 ➔ all data on the complete dataset
_t1 ➔ all data on a specific journal title
_t2 ➔ all data on a specific journal issue of title _t1
_t3 ➔ the first editorial item within the issue
_t3 ➔ the second editorial item within the issue
_t2 ➔ the second journal issue
_t3 ➔ the first editorial item within the issue
_t1 ➔ another journal title

Appendix. List of Site Visits during Planning Year

Date	Organization	Location	Purpose
26-30 March 2001	Elsevier Science National Library of the Netherlands	Amsterdam Den Haag	Fact-finding trip to learn about the production of electronic journals by ES and to learn about the digital archive work being done at the National Library of the Netherlands.
6-11 September 2001	British Library National Library of the Netherlands	London Den Haag	Validation of OAIS and OAI models to build prototype archives; learn about best practices from sites that have ongoing archival programs.
7 May 2001	J.P. Morgan Chase	Yale University New Haven, CT	Fact-finding visit to learn about potential economic benefits of outsourcing the storage component of an archive.
11-12 October 2001	Elsevier Science	Amsterdam	Fact-finding trip to learn more about potential content beyond traditional journals, the production process, and metadata population.

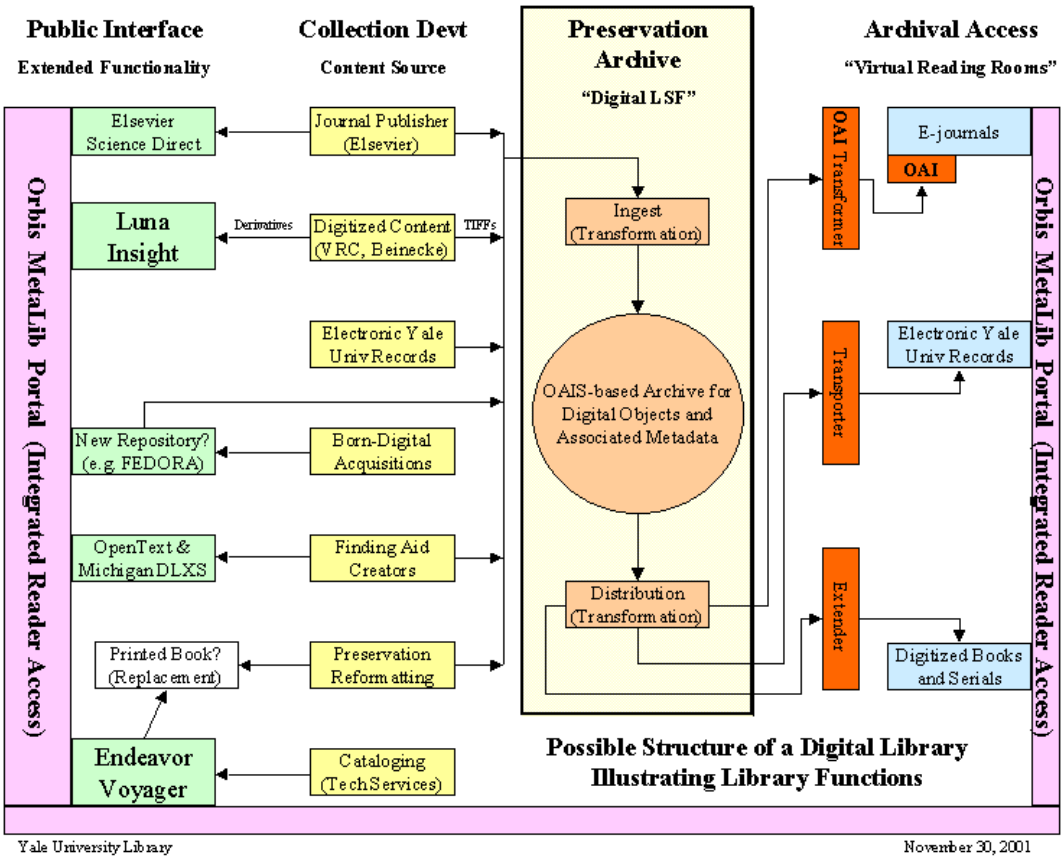
Appendix. Example of an XSLT stylesheet to transform MARC to DUBLIN CORE

Please see http://www.diglib.org/preserve/7A65.jpg

Appendix. Possible Structure for the Yale Digital Library

Desirable Characteristics of a Digital Library Infrastructure:

- Integration of system components
- Consolidation or aggregation of proliferating stand-alone databases
- Integration of a wide variety of digital objects and metadata schemas
- Integration of search interfaces and delivery mechanisms
- Flexible output: general, specialized, and personalized interfaces
- Interoperability with external systems and institutions
- Scalability
- Versatility
- Sophisticated management tools
- Direct focus on teaching and research needs



This diagram illustrates selected existing systems in the Yale Library and explores several future directions, with a focus on digital preservation. At the heart of the diagram is a new preservation archive for digital objects and associated metadata based on the Open Archival Information System model. The public interfaces on the right interact directly with this preservation archive. Those on the left rely upon completely independent systems where metadata and digital objects are stored separately from the archive. Content sources in the second column feed these systems in various ways:

Journal Publisher (Elsevier)

- Public access through full-featured online system maintained by vendor
- Formal partnership between Yale and vendor for archiving journal content
- Limited access to archive through OAI interface (Open Archives Initiative)

Digitized Content (Visual Resources, Beinecke, Digital Conversion Facility, Divinity, etc.)

- Public interface supplies sophisticated visual environment for teaching and study
- Insight system houses derived images (JPEGs and SIDs) and public metadata
- Archive houses original TIFF images and enhanced metadata
- Archive used only for image recovery or migration to new delivery platform

Electronic Yale University Records

- University records preserved for legal and historical purposes, low-use material
- Content sent directly to archive; no duplication of data in separate system
- Public interface retrieves digital objects directly from archive

Born-Digital Acquisitions

- Content is imported or directly input into new public repository
- Potential home for digital scholarship resulting from collaborative research projects
- Archival copies are transmitted from there to the archive

Finding Aids

- Finding aids distributed both to public service system and to archive

Preservation Reformatting (Digital Conversion)

- Digitized content sent directly to archive
- Hard-copy may be produced from digital version for public use
- Access to digital copy through custom application fed from archive
- Digital copy and original artifact may appear in national registry

Online Catalog

- Cataloging data resides only in LMS (NOTIS or Endeavor Voyager)

Integration achieved through MetaLib portal and lateral SFX links