

# **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, 4<sup>th</sup> Northumbria Conference, August 2001*

## Table of Contents

The Project Team.....	i
Acknowledgments.....	ii
Executive Summary.....	iii

### **PART I – Background, Approaches, Assumptions**

Challenges for Long-Term Electronic Archiving .....	2
Background for the Planning Project .....	7
Approaches and Assumptions.....	11
Plan of Work for the Year .....	18

### **PART II – Lines of Inquiry**

Triggering Events.....	22
Some Economic Considerations .....	25
Contract Between the Publisher and the Archive .....	31
Archival Uses of Electronic Scientific Journals.....	35
The Metadata Inquiry.....	40
Elsevier Science's Technical Systems and Processes.....	45
Creation of a Prototype Digital Archive.....	53
Digital Library Infrastructure at Yale.....	66

### **PART III – Appendices**

A: Three Models of Archival Agents .....	70
B: Dublin Core to Effect to SML Mapping (chart).....	73
C: Metadata Elements in an Electronic Journal Archive .....	74
D: Elsevier Science Technical Systems and Processes (definitions and flow charts) .....	102
E: OAIS Model: Overview Areas of Investigation in Mellon Project (diagrams) .....	117
F: List of Site Visits During Planning Year .....	125
G: Possible Structure of the Yale DLI (key and diagram) .....	126

## **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<sup>st</sup>, 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 proved 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 have donated far more of David Gewirtz's time than we had any right to expect and have offered their enthusiastic support and advice.

We have received help and insights along the way from the other Mellon planning projects and their staffs.

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 inter-dependence 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 differ from those required for production systems and add 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.

But 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 1200+ 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.