Changing Collaborations to Deliver Information in New Ways: Lessons Learned in the Illinois Digital Library Initiative Project

Timothy Cole and William Mischo

Introduction

Since the inception of the scientific journal in 1665, libraries and sci-tech publishers have enjoyed a synergistic and symbiotic relationship centered around the production and dissemination of the paper-based journal. In this traditional relationship, publishers have been responsible for acquiring content, packaging it, and distributing it. The role of the library has been to purchase the published product, organize it, make it accessible and then archive it for posterity. In FY '97 the ARL libraries purchased (across all disciplines) over $411 million in paper-format journals. The advent and growth of electronic journals, however, has begun to change this traditional relationship, and we can anticipate that the relationship will continue to evolve in the future.

The overall objectives—the acquisition, organization, dissemination, and preservation of information—haven't altered or become less important, but changes in journal format and delivery mechanisms are altering the ways in which libraries and publishers accomplish these objectives. As the information industry moves from paper-based to electronic journals and as publishers become involved in mounting their own complex information systems, the roles of publishers and libraries are changing. As users modify the way in which they access and process information, the responsibilities of those providing information services also change. This paper discusses our experiences with these evolving roles and responsibilities in the context of University of Illinois at Urbana-Champaign (UIUC) Digital Library Initiative (DLI) grant project. In this project we have explored issues connected with the local processing, loading, and indexing of electronic journals by libraries and mechanisms for extending the functionality of electronic journals.

The UIUC DLI Testbed

In the fall of 1994 UIUC was awarded a DLI grant to explore and develop new ways of delivering the full con-

Timothy Cole is systems librarian for digital projects and associate professor of library administration, and William Mischo is director of Grainger Engineering Library, University of Illinois at Urbana-Champaign.

tent of scholarly journals to users. In partnership with five leading publishers of scientific and technical journals and in consultation with a half-a-dozen other publishers in the field, the UIUC Library has built an online testbed presently containing approximately 60,000 full-content articles from 63 scholarly journals published since January 1995. The initial 4-year grant was funded jointly by the National Science Foundation, the Advanced Research Projects Agency, and the National Aeronautics and Space Administration. A follow-on 3-year grant funded by the Corporation for National Research Initiatives (CNRI) and the participating publishing partners began in the fall of 1998 at the expiration of the original grant. The goals of the ongoing UIUC DLI testbed project are:

- to create and make accessible to end-users a large-scale, multi-publisher, distributed repository of SGML/XML-formatted full-content journal literature in selected fields of science and engineering;
- to investigate issues relating to the local processing, indexing, normalization (including through use of metadata), retrieval, and rendering of journal literature in digital format;
- to study end user searching behavior and needs when using such materials online;
- to develop scalable models for effective dissemination and retrieval of information in an electronic full-content publishing environment;
- to evaluate methods for integrating full-content publishing resources with existing library information systems and services.

Key to the success of the project has been the identification and development of new relationships between the Library and journal publishers and between the Library and others, both on campus and off, involved in providing information services. The UIUC DLI project was created as a collaboration between the UIUC Library, the (then) National Center for Supercomputing Applications (NCSA), the UIUC Computer Science Department, and the UIUC Graduate School of Library and Information Science. Outside collaborators have included a core group of sci-tech publishers, researchers in the Computer Science Department at the University of Arizona, and a number of commercial computer software and hardware vendors.

Testbed Technologies

When work on the UIUC DLI project began, NCSA's Mosaic 2.0 beta was the browser of choice, the HTML 2.0 standard was still under development, Netscape had yet to release its first web browser, and Microsoft Windows 3.1 was the most common PC operating system. The initial task of the testbed project team was to identify technologies that were both of sufficient maturity to be usable at once and of sufficient potential to evolve over the life of the project.

First it was necessary to settle on a testbed document format standard. The ideal format would support full-text indexing; high-granularity, field-specific search and retrieval; and robust, platform-independent rendering. While no existing format matched all criteria and it was immediately obvious that HTML 2.0 fell far short of desired structure and rendering functionality, three other document format standards showed promise.

1. Standard Generalized Markup Language (SGML), a non-proprietary, international standard, was the best of the formats available in terms of exposing the intellectual structure of documents. The Text Encoding Initiative (TEI) was built around SGML and pilot, full-content journal publishing projects using SGML were then underway at OCLC. Rendering engines for SGML were weak, however.
2. TeX and LaTeX, were well established in the mathematical sciences academic community and supported extremely robust rendering of mathematics, but available authoring and viewing tools were limited and were largely UNIX-based. Exposure of document structure in TeX as used in real-world applications was limited.
3. PDF, an Adobe-proprietary format, provided the best emulation of the printed page. Adobe Acrobat reader was free and available for multiple platforms. However, PDF lacked (as of 1994) important hyperlink functionality and vital (for our project) cross-collection indexing features. It also was then, and remains today, a primarily appearance-oriented format.

In the end, SGML was chosen because it was non-proprietary and inherently best both for indexing and for search and retrieval. This was a decision made in consultation with perspective publishing partners. Though all had experience with the 3 formats under consideration, most favored SGML or SGML with embedded TeX for mathematical equations. The assumption was made that SGML rendering would catch up. To compensate for immediate SGML rendering limitations several of the publishers provided PDF versions of articles in conjunction with SGML versions.

Choosing an index/search engine was the next task. OpenText was chosen because it could exploit the strengths of SGML. (The OpenText search engine grew out of work done at the University of Waterloo to create and index the SGML version of the Oxford English Dictionary.) OpenText also had attractive features for indexing document metadata in conjunction with document full-text, for normalizing documents created with different publisher Document Type Definitions (DTDs), and for maintaining multiple, distributed repositories. Additionally, OpenText's architecture allowed us to integrate 3rd party tools, implement locally developed scripts and code, bypass OpenText modules we didn't need, and rapidly change processing procedures in response to dynamic research needs.

Originally the UIUC DLI project had expected to influence generic web client development by influencing development of the NCSA Mosaic web browser. This proved a naïve expectation. We quickly realized that search and delivery of testbed materials needed to be done in a browser-neutral manner. A focus of the testbed project has been the development of server-side scripts and CGI executables. For our servers, we've used both NT and UNIX operating system platforms as appropriate to task. Off-the-shelf webservers are used (initially Apache and the European Microsoft Windows NT Academic Center webservers, more recently Netscape Enterprise and Microsoft Internet Information Server webservers). Webserver functionality is extended using both conventional CGI and more advanced techniques such as Microsoft's Active Server Platform (ASP). HTTPS protocols (HTTP with Secure Socket Layers) are used for user authentication and authorization as required. HTTP protocols are used for all other interactions with clients.

Accomplishments to Date

By design the testbed is heterogeneous. Materials have been provided by five publishers: the American Institute of Physics (AIP); the American Physical Society (APS); the American Society of Civil Engineers; the Computer Society of the Institute of Electrical and Electronics Engineers (IEEE-CS); and the Institution of Electrical Engineers (IEE). These materials are created by various software systems and transmitted to us by various means (e.g., FTP, CD, and magnetic tape). Though all articles are provided in SGML, each publisher uses a different DTD (implying different tagging semantics). Materials are stored on multiple servers, and the indexes reside on different servers than those on which the articles reside. Our most significant accomplishment to date has been the demonstration of the viability of this distributed, heterogeneous repository model.

A variety of techniques have been used to accomplish this end result. Figure 1 shows the processing flow for the testbed. Materials are received from publishers and distributed to repository document servers. Pre-processing scripts are run which embed links to associated figures, check character entities, and extract and create a metadata file for each document (using RDF syntax and Dublin Core semantics supplemented with project-specific elements). Metadata is heavily used in the testbed both to normalize searching and to maintain link information between objects in the testbed and related objects external to the testbed. The project-specific metadata semantics go well beyond the minimal metadata tagging semantics of the Dublin Core and similar schema designed for general use on the Web.

OpenText indices are then built. Metadata is indexed along with full-text of the articles. Indices can be searched separately or in parallel. Tag aliasing is applied to support normalized searching. CGI and ASP scripts are used to enhance search functionality, insert hyperlink information, perform transformations between SGML, XML, and HTML, and facilitate linking between testbed objects and related information both within and external to the testbed.

Document repositories, metadata repositories, and the actual indices can be maintained on separate servers and can be searched separately or simultaneously. Figure 2 provides another view of the testbed architecture. Web browsers communicate (via HTTP & HTTPS) with CGI and ASP gateways that search the testbed indices. These indices were built from document and metadata repositories that may be collocated or distributed. Indices contain pointers both to testbed documents and to information resources in other systems. The gateways can filter links sent back to the client based on client characteristics and established authorization.

The system has been available to end users campus-wide at UIUC since October 1997, and at Notre Dame University for the past 6 months. 10,106 **end-user** search sessions have been conducted to date.

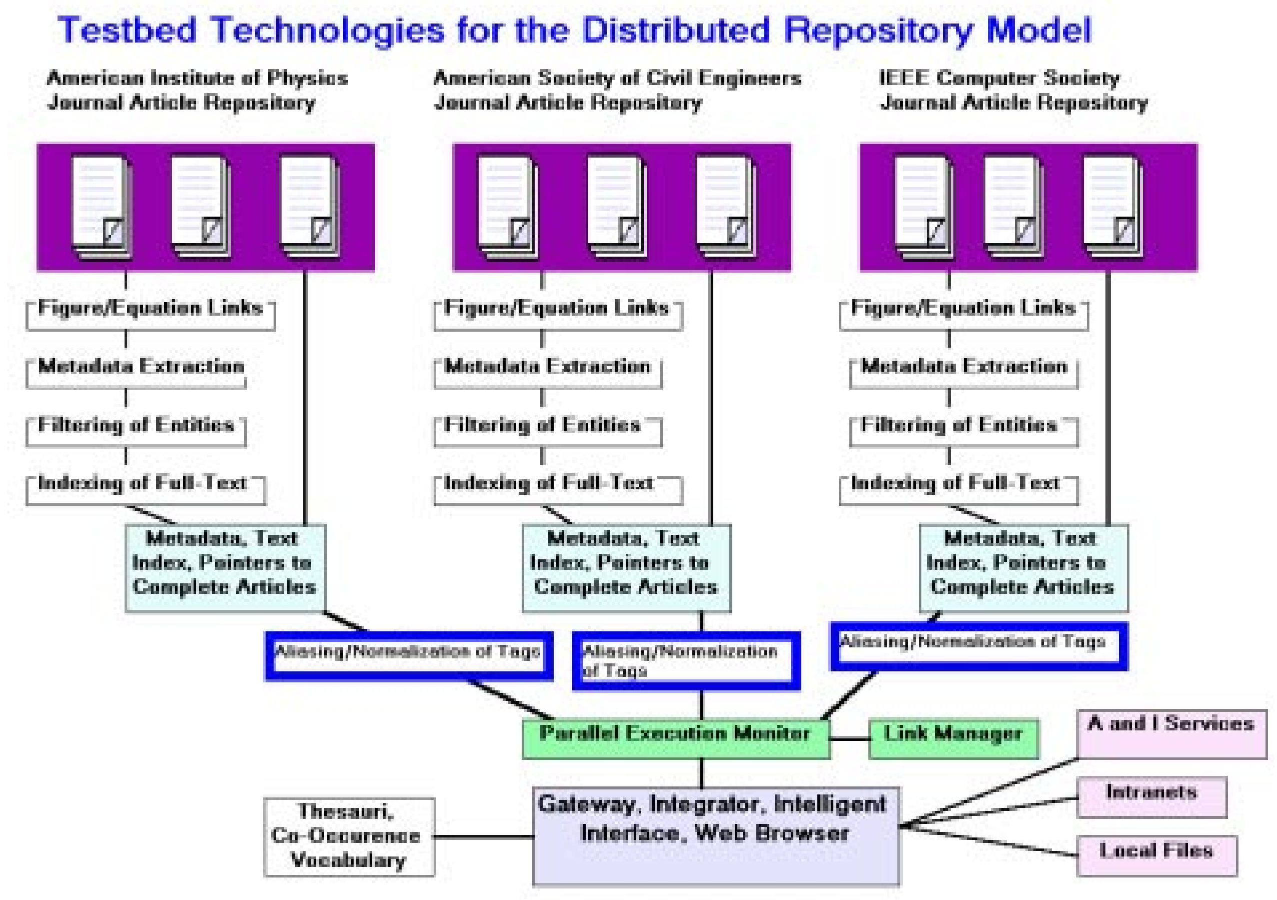

Testbed Technologies for the Distributed Repository Model
American Institute of Physics Journal Article Repository
American Society of Civil Engineers Journal Article Repository
IEEE Computer Society Journal Article Repository
Figure/Equation Links
Metadata Extraction
Filtering of Entities
Indexing of Full-Text
Metadata, Text Index, Pointers to Complete Articles
Aliasing/Normalization of Tags
Parallel Execution Monitor
Link Manager
A and I Services
Intranets
Local Files
Gateway, Integrator, Intelligent Interface, Web Browser
Thesauri, Co-Occurence Vocabulary

Figure 1: DLI Testbed Processing Flow

Lessons Learned

1. The potential of SGML (and now XML) has been borne out by experience. The full-text indices are extremely rich, supporting a measure of search precision unavailable in previous full-text search systems. Figure 3 shows the search fields available to end-users in the current interface. SGML has greatly facilitated extraction of metadata and insertion of hyperlinks to related resources within and external to the testbed.

2. Rendering of complex mathematical mark-up continues to be problematic. Until recently we relied solely on the Panorama SGML viewer originally marketed by SoftQuad. In spite of promises to improve the rendering engine, development has lagged (Panorama was recently sold to Interleaf) and there still isn't a version of Panorama for the Macintosh. Rapid development of XML, advances in the latest version of HTML, and development of Cascading Style Sheets are improving prospects for better rendering. Nonetheless, our experience with Panorama demonstrates the degree to which libraries and information providers are dependent on the commercial sector for essential technology.

3. A detailed transaction log analysis of 4,158 end-user search sessions has been conducted.

Several interesting results have been gleaned from the transaction logs. These include: there is very little use being made of either 'Help' or 'Quicktips' functions; browsing of tables of contents is being performed in 39% of the search sessions; full-text searching is the predominant search mode, but in 24% of the sessions users chose a specific field; full-text is displayed in more sessions (69%) than extended citations (19%); in 25% of the sessions, users do multi-concept searching; and an average of 4 full-text documents are viewed per session.

4. Overall development of the testbed has taken longer than anticipated. With some notable exceptions (e.g., the lack of a robust SGML viewer), the technology needed has been available by the time needed. The development of processing procedures, the normalization of DTDs, and the development and implementation of metadata semantics have taken longer than anticipated. Technology infrastructure changes happen much more quickly than process changes that involve changing how libraries and information providers do their jobs.

5. Implementation of a digital information resource requires tighter integration of the parties involved. Small changes by a publisher in tagging semantics can require corresponding changes in indexing scripts, metadata extraction procedures, and further downstream, style sheet design. Conversely changes in browser software or rendering client can necessitate changes in tagging and indexing. Because each of these tasks may be performed by a different agency, close, efficient working relationships are essential.

6. In the electronic journal environment, roles and responsibilities are more fluid. While documents may reside on a publisher's server, metadata may reside elsewhere (e.g., on an abstracting and indexing service's hardware). Different agencies may create different metadata for the same objects (e.g., using different controlled vocabularies). Libraries may implement their own gateways and portals, or may contract for such services with consortia or other 3rd parties. A single article may be found through different gateways, using different index and metadata providers, even if full content of the article itself still comes from a single publisher's server. Archival responsibilities may be distributed between libraries, consortia, and publishers.

7. In the rapidly evolving electronic journals environment, academic libraries will need to re-examine their collection development policies in terms of ownership vs. access, become more actively involved in institutional and consortial licensing agreements, and become more actively involved in campus networking, server, and workstation policies and technologies.

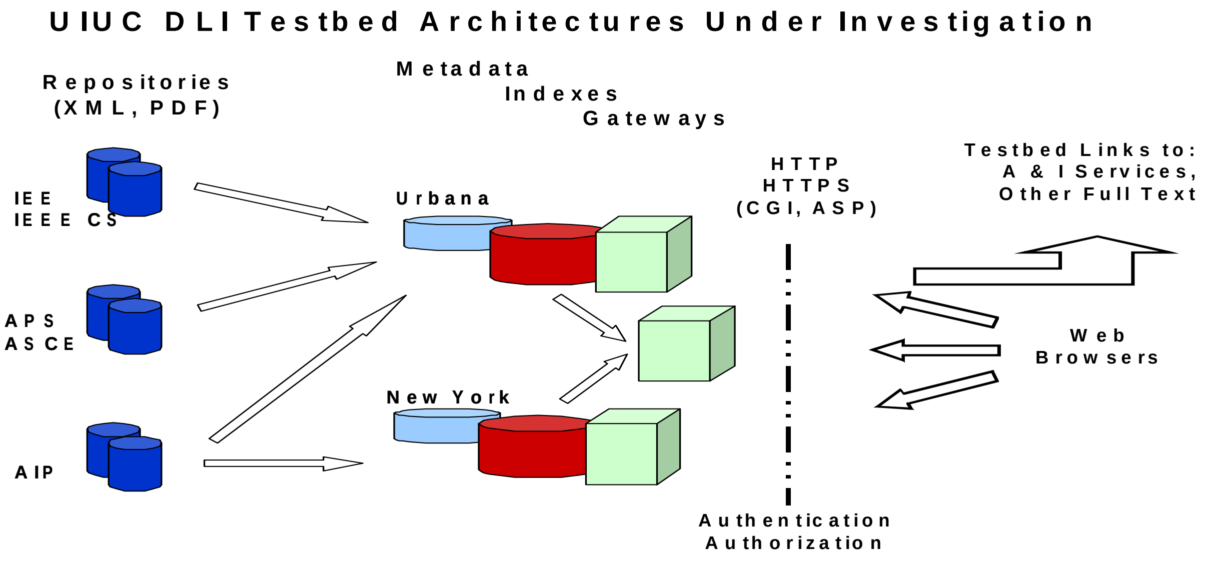
Future Foci

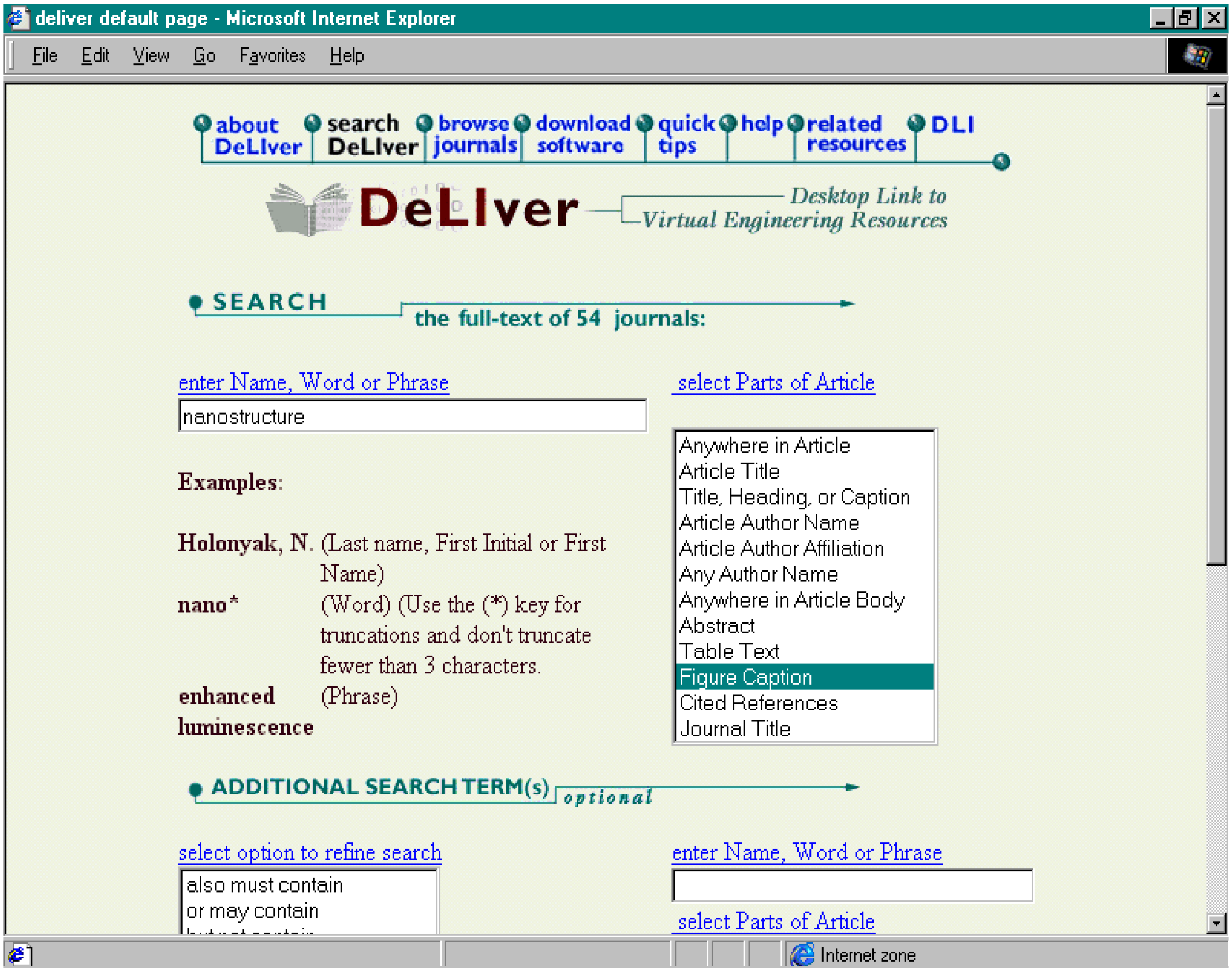
In the remaining years of this project additional issues and technologies will be investigated.

1. The entire testbed was recently converted to XML. Testbed articles are now retrievable in XML and HTML as well as in PDF and SGML. This has already improved rendering options and overall quality. The potential of the Math ML standard to support even better rendering of testbed content will be investigated.

2. Further testing of distributed architecture models will be done to test scalability and performance of the options.

Figure 2: UIUC DLI Testbed Architecture




deliver default page - Microsoft Internet Explorer
File Edit View Go Favorites Help
about DeLIver
search DeLIver
browse journals
download software
quick tips
help
related resources
DLI
DeLIver
Desktop Link to Virtual Engineering Resources
Internet zone

SEARCH the full-text of 54 journals:

enter Name, Word or Phrase

nanostructure

select Parts of Article

- Anywhere in Article
- Article Title
- Title, Heading, or Caption
- Article Author Name
- Article Author Affiliation
- Any Author Name
- Anywhere in Article Body
- Abstract
- Table Text
- Figure Caption
- Cited References
- Journal Title

Examples:

Holonyak, N.	(Last name, First Initial or First Name)
nano*	(Word) (Use the (*) key for truncations and don't truncate fewer than 3 characters.
enhanced luminescence	(Phrase)

ADDITIONAL SEARCH TERM(s) optional

select option to refine search

- also must contain
- or may contain
- but not contain

enter Name, Word or Phrase

select Parts of Article

Figure 3 — UIUC DLI Search Entry Screen

3. The use of Document Object Identifiers (DOIs) and other emerging standards to enhance and facilitate link management will be investigated.

4. Dynamic search mechanisms (e.g., word wheels to facilitate data entry, vocabulary switching to enhance search recall) will be integrated into our search interface.

5. Simultaneous search features (e.g., allowing simultaneous searching of non-testbed information resources) will be further tested.

6. As part of the CNRI Digital Library Testbed suite,[3] the UIUC DLI testbed repositories and indices will be opened up to additional researchers investigating ways in which digital libraries of the future may evolve.

Notes

1. Association for Research Libraries. 1997. *Descriptive Statistics for Academic Institutions* [online]. Available: http://fisher.lib.virginia.edu/newarl/dstat.html [February 28, 1999].

2. The original UIUC DLI grant project included considerable work in addition to the DLI testbed described in this paper. For overviews of the complete grant project, see: Schatz, B., H. Chen, W. Mischo, T. Cole, J. Hardin, and A. Bishop. 1996. "Federating Diverse Collections of Scientific Literature." *Computer* 29 (5), 28–36. Schatz, B. et al. 1999. "Federated Search of Scientific Literature. *Computer,*" 32 (2), 51–59.

3. Corporation for National Research Initiatives 1999. *D-Lib Test Suite* [online]. Available: http://

www.dlib.org/test-suite/overview.html [February 28, 1999].

4. Sperberg-McQueen, C. M. 1994. "The Text Encoding Initiative: Electronic Text Markup for Research." In B. Sutton (ed.), *Literary Texts in an Electronic Age: Scholarly Implications and Library Services, Papers Presented at the 1994 Clinic on Library Applications of Data Processing, April 10-12, 1994*. 35–56. Urbana, Ill.: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

5. Both the CORE project, in which OCLC was a partner, and the OCLC Electronic Journals Online project were making use of SGML as of 1994. For further information about these projects as they existed in 1994, see: Weibel, S. 1994. "The CORE Project: Technical Shakedown Phase and Preliminary User Studies." *OCLC Systems and Services,* 10 (2 and 3), 99–102. Dykhuis, R. 1994. "The Promise of Electronic Publishing: OCLC's Program." *Computers in Libraries,* 14 (10), 20–22.

6. Cole, T. and M. Kazmer. 1995. "SGML as a Component of the Digital Library." *Library Hi Tech,* 13 (4), 75–90.

7. Terry, D. 1991. "Sidebar 4: Open Text Corporation." In J. Price-Wilkin, "Text Files in Libraries: Present Foundations and Future Directions." *Library Hi Tech,* 9 (3), 7–44.

8. W3C: World Wide Web Consortium. 1999. *Resource Description Framework (RDF) Model and Syntax Specification: W3C Recommendation 22 February 1999* [online]. Available: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/ [February 28, 1999].

9. The Dublin Core Metadata Initiative. 1998. *The Dublin Core: A Simple Content Description Model for Electronic Resources* [online]. Available: http://purl.oclc.org/dc/ [February 28, 1999].

10. W3C: World Wide Web Consortium. 1998. *Mathematical Markup Language (Math ML) 1.0 Specification: W3C Recommendation 07-April-1998* [online]. Available: http://www.w3.org/TR/1998/REC-MathML-19980407/ [February 28, 1999].

11. Technology Update: Digital Object Identifiers. 1998. *Online & CD-ROM Review,* 22 (2), 115–18. See also: International DOI Foundation. 1998. *The Digital Object Identifier System* [online]. Available: http://www.doi.org/ [February 28, 1999].